COMPUTING IN CRYSTALLOGRAPHY

Editors:
R. DIAMOND
S. RAMASESHAN
K. VENKATESAN

Published by The Indian Academy of Sciences
and the International Union of Crystallography

HEAVY ATOM POSITIONS IN MACROMOLECULES

GOPINATH KARTHA

Roswell Park Memorial Institute
666 Elm Street, Buffalo, New York 14263

SUMMARY In the x-ray crystallographic structure analysis of macromolecules containing thousands of atoms, much success has been achieved during the last decade by the application of multiple isomorphous series and anomalous scattering methods. With these techniques the attempt is to obtain information regarding the missing relative phase angles of reflections by converting phase information into measurable amplitude information. This is achieved by measuring amplitude change that occurs to the structure factor vector when a scattering vector of known amplitude and phase is added. It is seen that simple relationships exist relating the changes in amplitude to the unknown plase angle. In practice this is achieved by preparation of heavy atom isomorphs of the crystal and measuring the diffracted intensities. A proper application of this method presupposes knowledge of the locations and scattering properties of the heavy atom substituents and it is seen that these parameters can also be obtained and refined with the use of the diffraction intensities themselves by the application of well established crystallographic techniques.

Introduction Developments in direct methods during the past decade has made the structure elucidation of medium sized molecules containing a fifty atoms or so a fairly routine matter. However, many macromolecules that are being studied at present by x-ray diffraction have molecular weights in the tens of thousands of Daltons. The complete description of their three dimensional structure involves the knowledge of the positions of a few thousand atoms. The experimental and computational difficulties are complicated by the fact that very few macromolecules give diffraction data beyond 2A° resolution. Even with the best available crystal, this lack of resolution makes the ratio of the number of observations to the number of atomic parameters to be determined to be of the order of two or three. Hence, at present the direct statistical methods of obtaining phase information which made use of the over abundance of observations compared to the number of atomic parameters that occurred in the small molecule crystallography has been of very limited use in macromolecular studies. The successful structure determination of these large molecules till now had depended on the use of other indirect information regarding certain known structural features in the crystal and their scattering behaviour to help in the application of the x-ray diffraction methods. In particular, the tremendous successes in the determination of globular protein structures during the last decade came through the use of the isomorphous replacement and anomalous scattering methods.

Principle of the method To build up a picture of the unit cell contents of the crystal one needs not only the measured structure amplitudes of the reflections but also their relative phases. The x-ray diffraction measurement, however, yields only the amplitude of the scattered wave but not its phase. Hence, any experimental strategy for obtaining the relative phases of the reflected waves are in essence methods of converting phase differences into measurable amplitude differences. The principle is illustrated in Fig. 1.

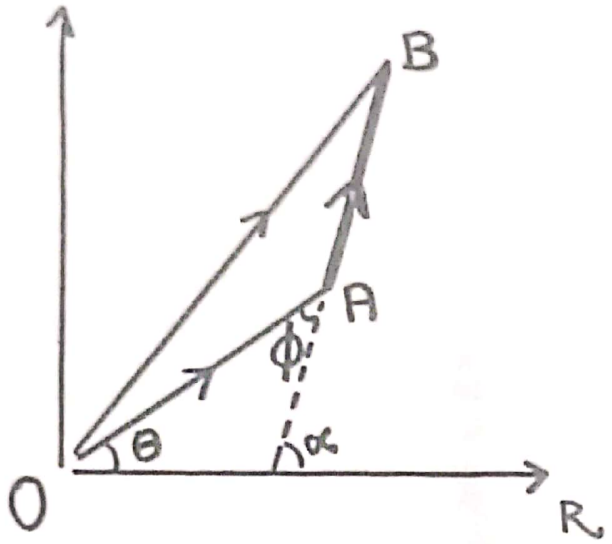


Fig. 1

Let us assume that the vector OA represents a structure factor of known amplitude but undetermined phase angle θ. In order to determine θ, we add to OA, a probe vector AB, whose amplitude as well as phase angle α are known. The resultant vector is OB the amplitude of which can also be experimentally measured. By using the probe AB we aim at obtaining information regarding the unknown phase angle θ of OA interms of the measured amplitudes OA and OB as well as the amplitude AB and phase angle α of the probe vector AB.

From the phase triangle OAB one can easily write down an expression for the phase angle difference between the unknown vector OA and the probe vector AB from only the magnitudes of the three vectors. We have $\text{Cos } \phi = -(OA^2 + AB^2 - OB^2)/2OA.AB$.....(1) and it is seen that the expression gives only the magnitude of the angle difference between OA and AB. In other words we do not know whether it is a phase advance or retardation as the phase triangle can be constructed equally well on either side of the probe vector AB.

Knowing the phase α of AB, the phase angle θ of OA can be either of the two values given by

$$\theta = \alpha \pm \phi \quad \ldots\ldots(2)$$

The above equation indicates that the amplitude change occurring on addition of a known vector gives information only about the component of the unknown vector in the direction of the probe vector. By probing $\underset{\sim}{OA}$ by two known but non-collinear vectors and observing the resulting amplitudes, one should be able to obtain the phase angle of $\underset{\sim}{OA}$ unambiguously.

This is the principle used in the experimental solution of the phase problem in macromolecules by the use of multiple isomorphous series[1] and anomalous scattering methods[2,3]. In use of the isomorphous series method, intensity data from the parent crystal as well as from isomorphous derivative crystals in which a few additional heavy atoms are bound at specific sites on the macromolecule are made. The probe vector here is then the scattering vector from the additional heavy atoms and is known if the location and scattering parameters of these atoms are also known. In the use of anomalous scattering techniques one exploits the fact that the heavy atom substituents in the derivative crystals are in general anomalous scatterers and for suitable wavelengths give a component which has a phase advance. The out of phase component results in the structure amplitudes $\underset{\sim}{F}(\underset{\sim}{h})$ and $\underset{\sim}{F}(\bar{\underset{\sim}{h}})$ of the Friedel pairs of reflections being unequal in the non-centrosymmetric crystals, a difference which, though small, can be experimentally measured. In the anomalous scattering method it is this out of phase component that acts as the probe vector and yields phase information from the Bijovet differences.

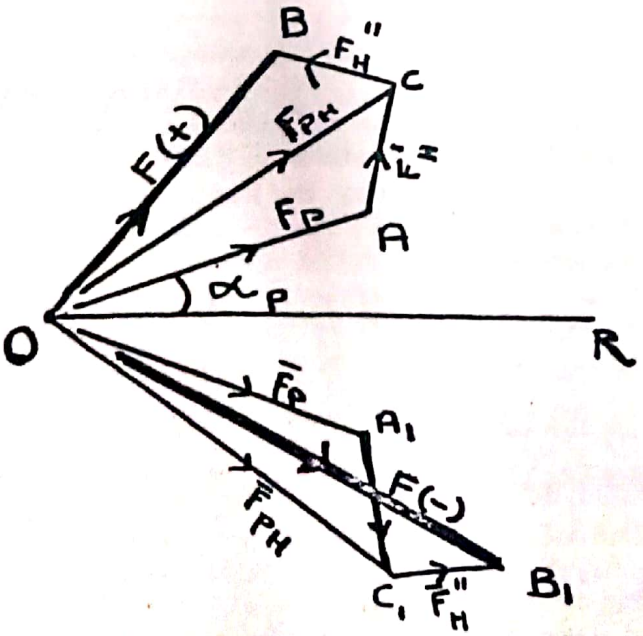


Fig. 2

The Fig. 2 shows the relationships between the various scattering vectors involved and their amplitude relationships are shown in Fig. 3 where the vector triangle has been reflected on the real axis.

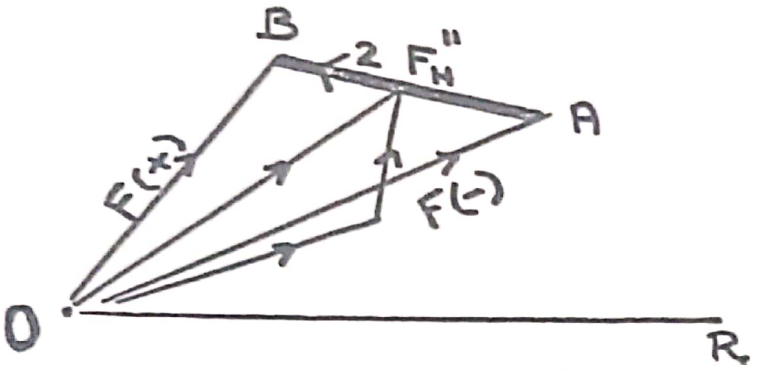


Fig. 3

Here also two solutions corresponding to construction of the phase triangles on either side of the probe vector are possible and are symmetrical with respect to it. However, unlike the isomorphous replacement case, the two solutions correspond to different values for the amplitude Fp of the native protein.

The evaluation of the protein phase angles and their refinement will be the subject matter of another lecture and hence will not be discussed here. However, it may be mentioned that in practice the determination of the heavy atom parameter and protein phase evaluation are so correlated that during any actual analysis these two steps proceed in a series of interleaved cycles; the phase angle knowledge at any stage affecting the accuracy of the heavy atom parameters and vice versa.

Location of heavy atoms.

Heavy atom scattering vector.

The first step[4] in applying the isomorphous series method is the determination of the nature and position of the heavy atoms in the derivative crystals. Once these parameters are known the probe vector $\underset{\sim}{F}_H$ can be calculated both in magnitude and direction. If these atoms are also anomalous scatterers then both the individual atomic scattering factors $\underset{\sim}{f}_j$ and the total heavy atom structure factor $\underset{\sim}{F}_H$ are both complex numbers. They can further be split into

a scattering component in phase with the incident beam and another out of phase component. Thus, we can write

$$\underset{\sim}{f}_j = f'_j + if''_j$$

and

where

$$\underset{\sim}{F}_H(\underset{\sim}{h}) = F'_H(\underset{\sim}{h}) + F''_H(\underset{\sim}{h}) \quad \ldots\ldots\ldots(3)$$

$$\underset{\sim}{F}_H(\underset{\sim}{h}) = \sum_j f'_j(\underset{\sim}{h}) \exp 2\pi i(\underset{\sim}{h}.\underset{\sim}{r}_j)$$
$$= F'_H e^{i\alpha'_H} \quad \ldots\ldots\ldots(4)$$

and

$$\underset{\sim}{F}''_H(\underset{\sim}{h}) = i\sum_j f''_j(\underset{\sim}{h}) \exp 2\pi i(\underset{\sim}{h}.\underset{\sim}{r}_j)$$
$$= F'' e^{i(\alpha'_H + \omega)}$$
$$= F''_H e^{i\alpha''_H} \quad \ldots\ldots\ldots(5)$$

In the above equations $\underset{\sim}{h}$ defines the reciprocal lattice vector for which the values are given, $\underset{\sim}{r}_j$ the position vector of atom j, f' and f" are the real and imaginary components of the atomic scattering factors. Knowing $\underset{\sim}{r}_j$ and $\underset{\sim}{f}_j$ of all substituent atoms, the phase angles α'_H and α''_H corresponding to the components F'_H and F''_H can be evaluated from equations 4 and 5. We shall now consider how the positions of the substituent atoms can be obtained.

The structure factors F_H, F_p and F_{PH} of the heavy atoms, the protein and the derivative are related by the vector equation.

$$\underset{\sim}{F}_H = \underset{\sim}{F}_{PH} - \underset{\sim}{F}_H = \underset{\sim}{F}_{PH} - \underset{\sim}{F}_p \quad \ldots\ldots\ldots(6)$$

The situation here is the converse of what is shown in Fig. 1. Here, the aim is to obtain the difference vector $\underset{\sim}{F}_H$ from the observed amplitudes of F_{PH} and F_p. From an estimate of F_H, we can obtain the nature and location of atoms which give rise to $\underset{\sim}{F}_H$ by Fourier methods. However, the magnitudes of F_{PH} and Fp at best gives only the length of $\underset{\sim}{F}_H$ and hence the atomic parameters themselves must be obtained from these by the application of Patterson or direct methods.

Length of $\underset{\sim}{F}_H$ From the vector triangle OAB of Fig. 1 we get

$$OB = OA \cos A\hat{O}B + AB \cos A\hat{B}O \quad \ldots\ldots(7)$$

as the length of vector $\underset{\sim}{OB}$. Considering the phase triangles occurring in the case of isomorphous series and anomalous scattering differences and neglecting small terms[5,6,7] we can write

$$F_{PH}(\underset{\sim}{h}) - F_p(\underset{\sim}{h}) \triangleq F'_H(\underset{\sim}{h}) \cos(\alpha_{PH} - \alpha'_H) \quad \ldots\ldots\ldots(7)$$

$$F_{PH}(\underset{\sim}{h}) - F_{PH}(\bar{\underset{\sim}{h}}) \triangleq 2F''_H(\underset{\sim}{h}) \cos(\alpha_{PH} - \alpha''_H) \quad \ldots\ldots\ldots(8)$$

which gives the measured amplitude differences in terms of the heavy atom parameters and phase angles of derivatives. Knowing all the atomic parameters, the above equations imply a complete solution of the phase angle α_{PH}.

Further simplification can be obtained if all the substituent

atoms have the same (f"/f') value which will result in the phase of the imaginary component being $\pi/2$ in advance of the real component resulting in

$$\alpha''_H = \alpha'_H + \pi/2$$
$$2\ F''_H = kF'_H$$

and leads to the amplitude differences expression

$$\Delta\ Fiso = F_{PH} - Fp = F'_H\ COS\ (\alpha_{PH} - \alpha'_H) \qquad (9)$$

and

$$\Delta\ Fano = \frac{1}{k}\ [F_{PH}(\underset{\sim}{h}) - F_{PH}(\bar{\underset{\sim}{h}})] = F'_H\ \sin(\alpha_{PH} - \alpha'_H)$$

we also get

$$F'_H = (\Delta\ F^2_{iso} + \Delta\ F^2_{ano})^{1/2}$$
$$\alpha_{PH} = COS^{-1}\ (\Delta\ F_{iso}/F'_H) + \alpha'_H \qquad (10)$$
$$= Sin^{-1}\ (\Delta\ F_{ano}/F'_H) + \alpha'_H$$

The heavy atom scattering vector can be represented in terms of the isomorphous series and anomalous scattering amplitude differences as

$$F_H = (\Delta\ F_{iso} - i\ \Delta\ F_{ano})\ exp\ i\alpha_{PH} \qquad .....(11)$$

from which it is seen that the heavy atom configuration may be obtained as a Fourier series from the measured differences if the phases α_{PH} are known. However, in the initial stages of the analysis no phase information may be available for this method of locating the heavy atoms.

Heavy atom vector maps. The amplitude differences Δ Fiso and Δ Fano may be either positive or negative for the different reflections but their magnitudes can be large only if the corresponding heavy atom vector F_H is also large. For centro symmetric projections the relationship between heavy atom scattering vector and the structure amplitudes are straight forward and hence the earlier studies were mainly restricted to this small class of reflections. However, if accurate anomalous scattering measurements are also available, then good estimates of the length of heavy atom vector F_H could be obtained by appropriate combination [6,8] of the differences. Vector maps computed with the square of the amplitude differences have been used by many investigators [9,10,11] and has become one of the first essential steps in any protein structure analysis. It can easily be seen from the simplified expression given in equation 9 that

$$|\Delta\ F_{iso}|^2 \simeq |F_H|^2\ \cos^2\ (\alpha_{PH} - \alpha_H)$$
$$\simeq 1/2\ |F_H|^2 + \tfrac{1}{2}|F_H|^2\ \cos\ 2(\alpha_{PH} - \alpha_P) \qquad .....(12)$$

Inspection of equation (12) shows that the first term on the right hand side is the appropriate coefficient for computing the Patterson map of the heavy atom structure and yields positive peaks of height

$1/2\ f_{Hi}\ f_{Hj}$ at locations $\pm(r_i - r_j)$. The second term, however, depends not only on the heavy atom configuration, but also on the protein density distribution and hence leads to a background of positive and negative peaks which tends to mask the features of the heavy atom vectors we seek to locate.

Similar heavy atom vector maps could be constructed from the anomalous scattering differences[12] which also can be shown to consist of heavy atom vector peaks superimposed on fluctua ting background of positive and negative peaks in a way anticomplementary with[5] those occurring in the isomorphous series map. Hence, appropriate combinations of the two differences are better in terms of cancelling unwanted background features but accentuating the required heavy atom vector peaks. In using the amplitude differences, it is essential to remember that these are in general small differences between two large measurements which are subjected to experimental errors and hence appropriate error analysis and weighting procedures are essential. Systematic errors due to lack of isomorphions, absorption of x-rays etc. should be carefully evaluated and sometimes local scaling [13] procedures may significantly improve the results. In general, the vector maps of anything but simple groupings of a few heavy atom substituents are difficult to analyse in an essentially complete manner.

In most macromolecular structures investigated so far, the calculation and inspection of vector maps to locate heavy atoms have been one of the first essential steps in the successful analysis. Both $(\Delta \text{ Fiso})^2$ and $(\Delta \text{ Fano})^2$ as well as suitable weighted combinations are usually examined. Some proteins contain heavy atoms like iron, which show appreciable anomalous scattering when irradiated by x-rays of ordinary wavelengths. In such cases it is possible to use $(\Delta \text{ Fano})^2$ type maps using the native protein data to locate these atoms by accurate measurement of Bijovet differences as was done in the case of calf liver cytochrome b_5.

Correlation of heavy atom positions. While the relative configuration of the substituent atoms in a single derivative can be obtained from vector maps, in many space groups their co-ordinates are not given with reference to a unique origin. For using the multiple isomorphous series for protein phase evaluation, it is essential that all the heavy atom co-ordinates in the different derivatives be referred to a common origin in the unit cell. In space groups like $P2_1$ or $P6_v$ where the choice of origin might involve a variable parameter also, additional complications may arise. Many vector methods have been proposed [5,10,12,15] to do this and here again the combination of isomorphous derivative and anomalous scattering data can greatly simplify the interpretation of the results.

However, if approximate initial protein phases α_p are available, say by use of single isomorphous series and anomalous scattering methods, then the locations of the substituent heavy atoms in the

above equation being chosen which gives the smaller difference between $|F_H|$obs and $|F_H|$calc. This method, however, in addition to using only a very small part of the data was also not able to refine some of the parameters, such as the y parameter in space group $P2_1$. Use of anomalous scattering information can give [6] a better estimate of $|F_H|$obs against which the heavy atom parameters could be refined from the expression

$$|F_H|_{obs}^2 = (\Delta|Fiso|)^2 + 2|F_{PH}||F_p|[1 - \{1 - (\Delta Fano|Fp)^2\}^{1/2}] \quad \ldots\ldots\ldots(14)$$

which gives the length of the heavy atom vector in terms of the measured quantities. For optimum use of the least squares method of refinement it is essential that proper estimate of errors be applied and appropriate weighting schemes used. In actual cases the R factors with the final parameters obtained after least squares refinement lie in the range of 25 to 45% and as such is considerably higher than occurring in small molecule studies. The actual agreement index R for any particular case depends on accuracy of data, numbers and types of heavy atoms, resolution of data, degree of isomorphism between native protein and derivative, molecular weight of protein etc.

Another refinement procedure [18] is to use the protein phase calculated from the remaining derivatives in defining a lack of closure.

$$\xi_j = F_{PH}(obs) - F_{PH}(calc)$$

and minimizing $\sum \omega_j^2$

the sum over all reflections by usual least squares methods. Here the weight to used for each reflection is $w = 1/(\sum_j \xi_j^2)^{1/2}$ i.e. the inverse of root mean square of the lack of closure over all derivatives used in calculating the protein phase angle: Different kinds of reliability indices have been suggested [19] and have been found helpful in following the course of parameter refinement.

Acknowledgement. I thank the National Institute of Health, National Science Foundation and New York State Department of Health for support.

Legend to Figures

Fig. 1 The probe vector $\underline{AB}$ of known magnitude and phase is added to unknown vector $\underline{OA}$ to produce the resultant $\underline{OB}$. Angle ϕ between vectors $\underline{OA}$ and $\underline{AB}$ can be obtained from the magnitudes of the three vectors.

Fig. 2 Relationship between the structure factor vectors of Friedel pairs of reflections. The amplitudes of $\underset{\sim}{F}(\underline{h})$ and $\underset{\sim}{F}(\bar{\underline{h}})$ are, in general, different if there are anomalous scatterers in the unit cell.

Fig. 3 Effect of out of phase component F_H'' on the structure amplitude $\underset{\sim}{F}(\underline{h})$ and $\underset{\sim}{F}(\bar{\underline{h}})$. Angle between $\underset{\sim}{F}(\underline{h})$ and $\underset{\sim}{F}_H''$ can be calculated from these amplitudes.

References

(1) Green, D.W., Ingram, V.M. and Perutz, M.F. 1954. Proc. Roy. Soc. (London) A225, 287.
(2) Okaya, Y., Saito, Y. and Pepinsky, R. 1955. Phys. Rev. 98, 185.
(3) Ramachandran, G.N. and Raman, S. 1956. Curr. Sci. 25, 348.
(4) Blundell, T.L. and Johnson, L.N. 1976. Protein Crystallography, Academic Press, London.
(5) Kartha, G. and Parthasarathy, R. 1965. Acta Cryst. 18, 745, 749.
(6) Kartha, G. 1965. Acta Cryst. 19, 883.
(7) Kartha, G. 1975. Anomalous Scattering, Ed. S. Ramaseshan and S.C. Abrahams, Munksgaard, Copenhagen, 363.
(8) Singh, A.K and Ramaseshan, S. 1966. Acta Cryst 21, 279.
(9) Perutz, M.F. 1956. Acta Cryst 9, 867.
(10) Kartha, G., Bello, J., Harker, D. and DeJarnette, F.E. 1963. Aspects of Protein Structure, Ed. G.N. Ramachandran, New York, Academic Press, P. 13.
(11) Mathews, B.W., Fenna, R.E., Bolognesi, M.C., Schmid, M.F. and Olson, J.M. 1979. J. Mol. Biol. 131, 259.
(12) Rossman, M.G. 1960. Acta Cryst. 13, 221.
(13) Mathews, B.W. and Czerwinski, E.W. 1975. Acta Cryst. A 31, 480.
(14) Mathews, F.S., Levine, M. and Argos, P. 1972. J. Mol. Biol. 64, 449.
(15) Steinrauf, L.K. 1963. Acta Cryst. 16, 317.
(16) Dodson, E. and Vijayan, M. 1971. Acta Cryst. B 27, 2402.
(17) Hart, R.G. 1961. Acta Cryst. 16, 1196.
(18) Dickerson, R.E., Kendrew, J.C. and Strandberg, B.E. 1961. Acta Cryst. 14, 1188.
(19) Kraut, J., Sieker, L.C., High, D.F. and Freer, S.T. P.N.A.S. U.S.A. 1962, 48, 1417.

EXERCISE

The following diagrams represent sections from Patterson maps for a ribonuclease A platinum derivative to a resolution of 3Å. The relevant information is given below:

a = 30.13	space group $P2_1$
b = 38.11	
c = 53.29	Z = 2
β = 105.75	

The coefficients used for calculating the maps are

a, $(|F_{pH}|-|F_p|)^2 = A$

b, $(|F_{pH}|-|F_{pH}|)^2 = B$

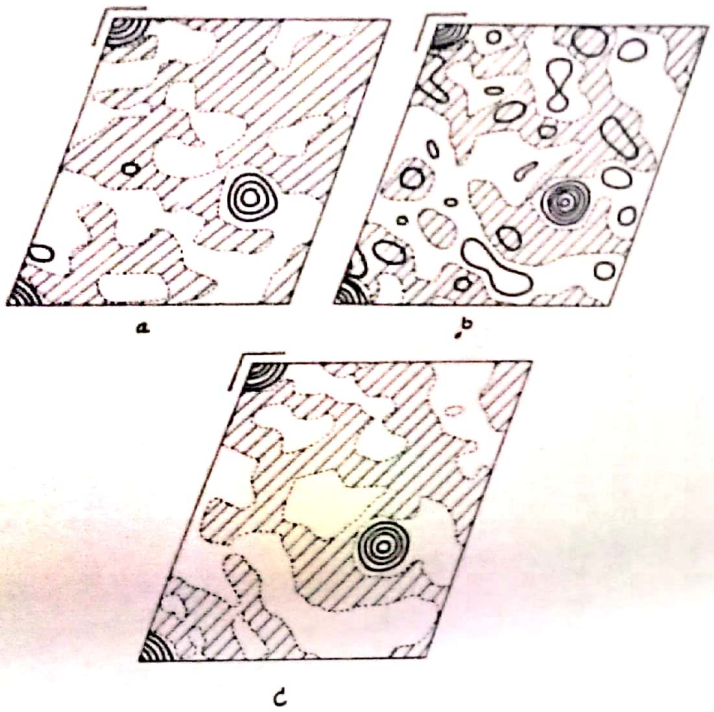
c, $W_1 \times A + W_2 \times B$ where W_1 and W_2 are weights.

The maps are drawn as sections perpendicular to b-axis for

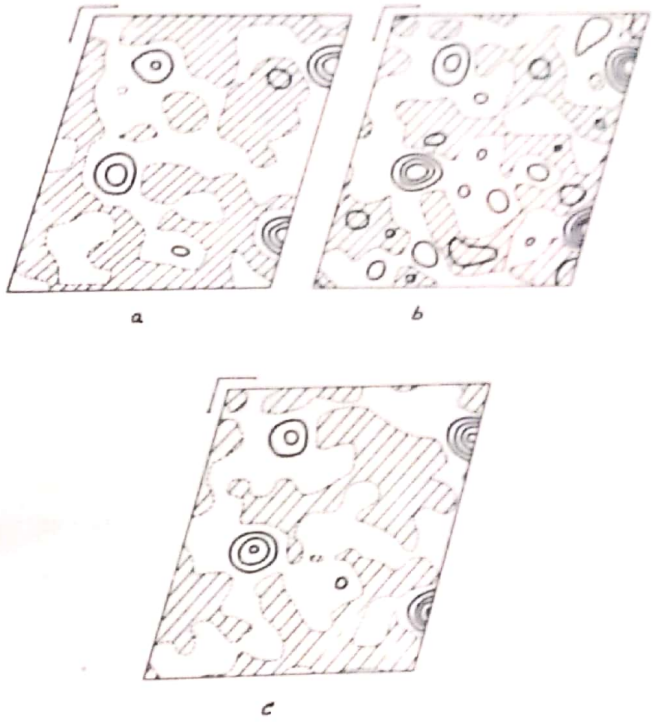
o —— c/2 horizontal
o —— a down

Two sets of maps corresponding to Y = o and Y = 1/2 show all main peaks in the maps.

Find the number and locations of the heavy atom substituents.



Section y = 0



Section y = 1/2

G. KARTHA

R NASE. CDG. DIFF. PATTERSON SECTIONS

THE REFINEMENT OF CRYSTAL STRUCTURES BY FAST-FOURIER LEAST-SQUARES

RAMESH C. AGARWAL

Centre for Applied Research in Electronics
Indian Institute of Technology, New Delhi 110 029, India

Summary

A least-squares atomic parameter refinement method is described which makes use of the Fast-Fourier transform (FFT) algorithm at all stages of the computation. Therefore, the computational requirement is proportional to NLogN, where N is the number of reflections, making it very attractive for large structures such as proteins. The method has a radius of convergence of approximately 0.75 Å, making it attractive for a small structure also. Computational and programming considerations are described. Results of using the method on several structures are summarised.

Introduction

Recent results on the refinement of protein crystal structures with high resolution data (Huber, Kukla, Bode, Schwager, Bartels, Deisenhofer & Steigemann, 1974; Freer, Alden, Carter & Kraut, 1975; Moews & Kretsinger, 1975; Adman, Sieker & Jensen, 1975; Deisenhofer & Steigemann, 1975; Bode & Schwager, 1975; Chambers & Stroud, 1977; Takano 1977; Isaacs & Agarwal, 1978) have shown that refinement markedly improves the accuracy of the structure and the quality of the electron density map. With the exception of rubredoxin (Watenpaugh, Sieker, Herriott & Jensen, 1973) all the refinements listed here prior to 1978 were performed using either the real space method of Diamond (1971, 1974) or difference Fourier methods (see for example Watenpaugh et al. 1973). Rubredoxin has been extensively refined with difference-Fourier methods followed by block-diagonal least-squares refinement using a conventional program. The quality of this result, and the extra structural information (particularly the water structure) which result from this refinement showed the validity and value of refining protein structures by least-squares methods.

The principal obstacles to the routine use of least-squares refinement for protein structures are (1) the paucity of diffraction data relative to that for small molecules, and (2) the enormous computing cost.

THE REFINEMENT OF CRYSTAL STRUCTURES BY FAST-FOURIER LEAST-SQUARES

RAMESH C. AGARWAL

Centre for Applied Research in Electronics
Indian Institute of Technology, New Delhi 110 029, India

Summary

A least-squares atomic parameter refinement method is described which makes use of the Fast-Fourier transform (FFT) algorithm at all stages of the computation. Therefore, the computational requirement is proportional to NLogN, where N is the number of reflections, making it very attractive for large structures such as proteins. The method has a radius of convergence of approximately 0.75 Å, making it attractive for a small structure also. Computational and programming considerations are described. Results of using the method on several structures are summarised.

Introduction

Recent results on the refinement of protein crystal structures with high resolution data (Huber, Kukla, Bode, Schwager, Bartels, Deisenhofer & Steigemann, 1974; Freer, Alden, Carter & Kraut, 1975; Moews & Kretsinger, 1975; Adman, Sieker & Jensen, 1975; Deisenhofer & Steigemann, 1975; Bode & Schwager, 1975; Chambers & Stroud, 1977; Takano 1977; Isaacs & Agarwal, 1978) have shown that refinement markedly improves the accuracy of the structure and the quality of the electron density map. With the exception of rubredoxin (Watenpaugh, Sieker, Herriott & Jensen, 1973) all the refinements listed here prior to 1978 were performed using either the real space method of Diamond (1971, 1974) or difference Fourier methods (see for example Watenpaugh et al. 1973). Rubredoxin has been extensively refined with difference-Fourier methods followed by block-diagonal least-squares refinement using a conventional program. The quality of this result, and the extra structural information (particularly the water structure) which result from this refinement showed the validity and value of refining protein structures by least-squares methods.

The principal obstacles to the routine use of least-squares refinement for protein structures are (1) the paucity of diffraction data relative to that for small molecules, and (2) the enormous computing cost.

The paucity of diffraction data is characteristic of most protein crystals. The intensity of the diffracted radiation is reduced because of the large unit cell volume, and the extent of the scattering is further reduced by the large amount of disordered solvent in the cell. Few protein crystals have had data measured to a resolution of 1.5 Å. The reliability and accuracy of least-squares refinement reduces as the ratio of the number of observations to the number of parameters reduces. For practical purposes this means that diffraction data to a resolution of $\geq$ 2 Å is required for a meaningful refinement. Low temperature studies may be able to give high resolution diffraction data for proteins. The computing cost arises from the number of calculations required for least-squares refinement. For a full matrix least-squares calculation the computing required is proportional to NM^2 where N is the number of reflections and M the number of parameters. For the simplest diagonal least-squares calculation, the requirement is proportional to NM since a derivative of each calculated structure factor has to be computed for each variable parameter.

It is possible to circumvent a lack of data by either reducing the number of parameters or by adding additional observations into the calculation. The number of parameters may be conveniently reduced by treating groups of atoms as "rigid bodies" with their positions described by three rotational and three translational parameters. Alternatively, the number of observations may be increased by including information in the form of constraints on the known geometry (bond lengths and valence angles) of peptides. Least-squares programs have been written which use one (Konnert 1976) or both (Sussman, Holbrook, Chruch & Kim, 1977) of these approaches, or use least-squares coupled with potential energy minimization (Jack & Levitt, 1978). They have been used with remarkable success, particularly where only low resolution data is available, but both the methods using constraints or restraints are affected (although less acutely when rigid body constraints are used) by the rapid increase in the cost of computing with the increase in the size of the structure (Schmidt, Girling & Amma, 1977). The method of Jack and Levitt (1978) uses fast-Fourier least-squares procedures (Agarwal 1978).

This problem of cost required a new approach to least-squares refinement, and this was provided by Agarwal (1978), who showed that most of the calculations involved could be computed by Fourier transforms. With the availability of fast-Fourier transforms, FFT, (Cooley and Tukey 1965; Ten Eyck 1973; Winograd 1978) the algorithm is extremely fast, and the computing required is proportional to N log N where N is the number of reflections.

All the details of the method and results of its testing on several structures are contained in Agarwal (1978). Details of its application to a large protein are contained in Isaacs and Agarwal (1978). Although the method is most useful for large structures, it is applicable to small structures also because of its large radius of convergence (0.75 Å) and reduced computational requirement. Results of its application to a small structure are discussed in Agarwal (1978). Since the method has been discussed in detail earlier, we will confine ourself to a tutorial discussion.

The Method

In the least-squares refinement of atomic parameters the function minimized is

$$P = \frac{1}{2} \sum_{hkl} W_{(hkl)} [\,|F_{c(hkl)}| - |F_{obs(hkl)}|\,]^2$$

where W_{hkl} is a weighting function. This function is to be minimized with reference to atomic parameters. The corrections to the parameters are obtained from the matrix equation

$$\Delta u = -H^{-1} G$$

where Δu_i is the correction to be applied to the i^{th} parameter
H^{-1} is the inverse of the normal matrix whose general term is

$$H_{ij} = \sum_{r=1}^{N} W_r \frac{\partial |F_c(r)|}{\partial p_i} \frac{\partial |F_c(r)|}{\partial p_j}$$

where N is the number of reflections and W_r is a weighting function. G is the gradient vector (derivatives) of general form

$$G_i = \sum_{r=1}^{N} W_r(\Delta F(r)) \frac{\partial |F_c(r)|}{\partial p_i}$$

The size of the normal matrix is M x M where M is the number of parameters and the length of the gradient vector is M. The calculation of the gradient vector is proportional to NM and that of the normal matrix is proportional to NM^2.

There are three major computational steps in the refinement procedure. These are calculation of structure factors, the gradient vector, the normal matrix and its inverse. We briefly discuss how these can be calculated using FFT.

Calculation of Structure Factors

Structure factors are calculated by a FFT of a model electron density. This is not a new procedure. The use of the basic method was discussed by Sayre (1951) nearly thirty years ago. However, it is only with the development of FFT that the method has become viable for protein structures. Computationally there are two distinct stages in calculating structure factors. The first stage is to calculate the atom electron density of the structure at each point on a uniform grid parallel to the cell axes. The second is the Fourier inversion of this electron density map to obtain the structure factor magnitudes and phases. The computation required for the first stage depends on the fineness of the sampling interval throughout the cell, or the number of grid points, and the distance from the centre of each atom for which the electron density is to be computed.

The most expensive part of the structure factor calculation is setting up the model atom electron density which is the Fourier transform of the atom scattering factor curve corrected for thermal motion. The isotropic thermal motion of the atoms is represented as a Gaussian function $\exp(-B_m s^2/4)$ where s is $2 \sin\theta/\lambda$. If the atom scattering factor curve is also defined as a Gaussian function then the product of these two Gaussian functions is another Gaussian function whose Fourier transform is also Gaussian. Both Ten Eyck (1977) and Agarwal (1978) have given the formulae to calculate the atom electron density. The speed of this computation depends on the number of terms in the Gaussian approximation to the atom scattering factor curve and on the number of grid points for which the density is to be calculated for each atom. Coefficients for analytical approximations of four Gaussian terms are given in International Tables Vol. IV (1974). These give an accurate fit to the scattering factor curve over a wide range of $\sin\theta/\lambda$, but for most protein structures the range of angle is smaller and the curves can be approximated by a single Gaussian function, for data limited to low resolution, or by the sum of two or three Gaussian terms for higher resolution data (Agarwal 1978; Ten Eyck 1977). Reducing the radius of the atom will exclude a fraction of its density from the calculation. The fraction excluded beyond a given radius depends on the atom type and its temperature factor. Agarwal (1978) has computed this fraction for different limiting radii and temperature factors so that a radius consistent with the required accuracy of the calculation may be chosen.

Calculation of the Gradient Vector

Agarwal (1978) has derived the following expression for the gradient vector with respect to the x coordinate of the m^{th} atom

$$G(x_m) = \sum_s g_m(s)(-i2\pi h)W(s)E(s)\exp(i\phi(s))\exp(-i2\pi s.r_m)$$

where

$g_m(s) = f_m(s)\exp(-B_m s^2/4)$ = contribution of m^{th} atom to structure factors

$(s) = 2\sin\theta/\lambda$

$W(s) = W_{(hkl)}$ a weighting function

$E(s) = |F_{calc(hkl)}| - |F_{obs(hkl)}|$

$s.r_m = hx_m + ky_m + lz_m$

$\phi(s)$ = phase of $F_{calc(hkl)}$

Similar expressions hold for $G(y_m)$, $G(z_m)$ and $G(B_m)$ with the term $(-i2\pi h)$ replaced by $(-i2\pi k)$, $(-s^2/4)$ respectively.

$G(x_m)$ may be rewritten as

$$G(x_m) = \sum_s D_x(s)\, g_m(s)\exp(-i2\pi s.r_m)$$

where $D_x(s) = (-i2\pi h)\, W(s)E(s)\exp(i\phi(s))$

$G(x_m)$ then, is the Fourier transform of the product of two functions $D_x(s)$ and $g_m(s)$ evaluated at r_m (the position of the m^{th} atom). According to the convolution theorem, multiplication in reciprocal space is equivalent to convolution in real space. The Fourier transform of $g_m(s)$ is the electron density of the atom $\rho_m(r)$ and the Fourier transform of $D_x(s)$, which we shall call $d_x(r)$, is a modified difference density map. The gradient then is computed by the summation

$$G(x_m) = \sum_r d_x(r)\,\rho_m(r-r_m)$$

The computation of all the x derivatives requires the calculation of the modified difference density map, $d_x(r)$, by FFT followed by an integration of $d_x(r)$ with the electron density function for each atom. If the atom electron density is assumed to be zero outside a radius rad_m from the atom centre r_m, the summation need only be carried out within this radius for each atom. Separate difference density functions have to be computed for gradients with reference to y, z and B.

Calculation of the Normal Matrix

The following expressions for the normal matrix term $H(x_m, x_n)$, corresponding to interactions between x_m and x_n have been derived.

$$H(x_m, x_n) = H_1(x_m, x_n) + H_2(x_m, x_n)$$

where

$$H_1(x_m, x_n) = \sum_s \frac{1}{2} g_m(s)\, g_n(s)(4\pi^2 h^2) W(s) \exp(i2\pi s.(r_m - r_n))$$

$$H_2(x_m, x_n) = \sum_s -\frac{1}{2} g_m(s) g_n(s)(4\pi^2 h^2) W(s) \exp(i2\phi(s)) \exp(-i2\pi s. (r_m + r_n))$$

Similar expressions hold for all other elements inthe normal matrix, differing only in that the term $(4\pi^2 h^2)$ is replaced by a similar term depending on the type of interaction. For example, it is replaced by $(4\pi^2 k^2)$ for $y_m y_n$ interaction, by $(4\pi^2 hk)$ for $x_m y_n$ interactions, $(s^4/16)$ for $B_m B_n$ interactions, and $(i\pi h s^2)$ for $x_m B_n$ interactions.

If $A_{xx}(s) = 2\pi^2 h^2 W(s)$, then the $H_1(x_m, x_n)$ terms represent the Fourier transform of $A_{xx}(s)\, g_m(s) g_n(s)$ evaluated at $(r_n - r_m)$, the vector between the two atoms. $A_{xx}(s)\, g_m(s)\, g_n(s)$ is always real and positive. Its Fourier transform has a large peak at origin corresponding to the diagonal terms, then drops rapidly and alternates in sign as the distance between the atom increases. The $H_2(x_m, x_n)$ terms represent the Fourier transform of $-A_{xx}(s)\, g_m(s)\, g_n(s) \exp(i2\phi(s))$ evaluated at $(r_m + r_n)$. This involves phase terms so that, unlike $H_1(X_m, X_n)$, these terms will have no major peaks and their magnitude distribution is likely to be the same in all parts of the normal matrix. As the major contribution to the elements of the normal matrix comes from the H_1 terms, neglecting the H_2 contribution will not affect the final result, but only the rate of convergence.

For a diagonal least-squares approximation, m = n and

$$H_1(x_m, x_n) = \sum_s A_{xx}(s)\, g_m^2(s)$$

This may be computed directly following the procedure described in Agarwal (1978). The computation required is proportional to the number of unique reflections.

Off diagonal elements can be calculated in a similar manner to the gradients. We may write

$$H_1(x_m, x_n) = \sum_s A_{xx}(s)\, g_m(s)\, g_n(s) \exp(-i2\pi s.(r_n - r_m))$$

which is similar to the expression for the gradients except that $D_x(s)$ is replaced by $A_{xx}(s)$, $g_m(s)$ is replaced by $g_m(s)\, g_n(s)$ and r_m is replaced by $r_n - r_m$. If $a_{xx}(r)$ is the Fourier transform of $A_{xx}(s)$, and $\rho_{mn}(r)$,

the joint Gaussian electron density function of the m^{th} and n^{th} atoms, which is the Fourier transform of $g_n(s)\,g_m(s)$, then by the convolution theorem $H_1(x_m,x_n)$ is the convolution of $a_{xx}(r)$ and $\rho_{mn}(r)$ evaluated at $r=r_n-r_m$. This may be expressed as the summation:

$$H_1(x_m, x_n) = \sum_r a_{xx}(r)\,\rho_{mn}(r-r_n+r_m)$$

The summation is over all the grid points in real space, and $(r-r_n+r_m)$ is the distance of the grid point from the point (r_n-r_m). If the joint electron density $\rho_{mn}(r-r_n+r_m)$ is assumed to be non zero only for grid points within some limiting radius from the point (r_n-r_m), then the summation is much simplified. Furthermore, if the off-diagonal terms of the matrix are restricted to interactions between closely related atoms (atoms in the same side chain or the same peptide unit, for instance), then $a_{xx}(r)$ is required over a limited volume of real space about the origin and could be computed directly. Since $A_{xx}(s)$ will change only if the weights change, the function $a_{xx}(r)$ may be used in a number of refinement cycles until this happens. Similarly, the joint Gaussian electron density function $\rho_{mn}(r-r_n+r_m)$ will change only if (r_n-r_m), B_m or B_n changes. The calculation of other off diagonal terms is similar with A_{xx} replaced by the appropriate function as given above.

Programming Considerations

Although the algorithm may appear complex, writing a program is relatively straightforward. Agarwal (1978) has described all the procedures. The two largest computations are the fast-Fourier transforms and the modeling of the atom electron density. In all of these calculations it is important that the minimum amount of computation is done, not only to save computer time but also to save on storage.

(a) The Fast-Fourier Transforms

The nature of FFT means that the transforms are calculated for a complete unit cell, which requires a full set of data. Ten Eyck (1973) has shown, however, that symmetry may be utilised to reduce both the amount of computation and the amount of data required for the FFT and has written an excellent package of programs to perform these calculations. Savings in time and space may also be made when there are systematic absences in general (hkl) reflections.

(b) Modeling the Atom Electron Density

This is the most expensive part of the calculation and it is important that care is taken to optimise the programming. The use of single

and double Gaussian approxi mations to the atom electron density has been discussed by Ten Eyck (1977) and Agarwal (1978).

The only difficulty in the programming is to allow correctly for atoms which lie close to the edges of the asymmetric cell unit. This may be approached in two ways. In the first method, which was employed in the original program used for the insulin refinement (space group R3) the coordinates of each atom are transformed by the space group symmetry to lie within the cell asymmetric unit required for the transforms. If an atom extends outside this asymmetric unit, then this density is added to the symmetrically equivalent point within the asymmetric unit. This method is appropriate for a computer with a large virtual memory where the whole asymmetric unit may be considered to be held in core. In the case where only a small slab of density can be held in core this method is not efficient, since each atom in the structure has to be moved through each of the symmetry operators to determine if any part of its volume falls within the required slab. In the second procedure the initial atom coordinates are transformed through all the symmetry positions and a sorted list of those atoms which will have some density within the required asymmetric unit is retained. For any slab of density only those atoms contributing to the slab are used. The disadvantage of this system is that the atom list is very much extended and the exponential factors for duplicate atoms need to be computed a number of times.

The calculation of the gradients uses a similar routine except that premultiplication of the $W\Delta E$ values by -ih, -ik, -il may change the symmetry of the modified difference map. In R3 for example, premultiplication by -ih or -ik destroys the three fold symmetry around the origin which means that in convoluting the atom electron density with the modified difference map, special care has to be taken with atoms which extend over the edge of the asymmetric unit. This problem was overcome by using an expanded asymmetric unit for the modified difference map, which extended in both positive and negative directions on x and y by a distance greater than the maximum atom radius. This is a cumbersome procedure as it requires additional computer core to hold the section of the map and does not lend itself to producing a space group general routine. An alternative solution adopted by Eleonor Dodson (Baker & Dodson, 1979) is to use a symmetry expanded set of atoms, as for the structure factor calculation, and to compute the convolution product of the atom density and modified difference density only for that portion of the atom which lies within the required unit of the cell. If an atom has been shifted from its original position by a symmetry operation, the gradients are transformed through the inverse operation and the total atom gradients are summed from the individual contributions. In this way the calculation is space group general, and Dodson has written

programs for use in the space groups P_1, $P2_12_12_1$, $P4_12_12$, R3, $P3_221$.

The Requirements of the Method

In order to use the FFT least-squares certain conditions with regard to the resolution of the data set, the accuracy of the starting coordinates and the size of the computer have to be met. The accuracy of the final refined structure depends on the resolution of the data used, the higher the resolution the more accurate will be the structure. Generally, diffraction data to a resolution of at least 2 Å is required for a meaningful refinement. However, at the beginning of refinement, when the coordinate errors are large, high resolution terms should not be used and a refinement with data to a resolution of less than 2 Å may produce some improvement in the model structure.

Test calculations (Agarwal 1978) have indicated that the method is capable of correcting coordinate errors with an rms value of 0.75 Å. Obviously, a protein structure with this degree of error in the coordinates would not have the geometry expected for peptide units and it is likely that most model structures fitted to electron density maps will have smaller rms errors than this, although some individual atoms could have much larger errors. In both insulin (Isaacs & Agarwal, 1978) and actinidin (Baker & Dodson, 1979) the refinement was able to correct automatically coordinates which were in error by 0.5 Å on average. In actinidin, the starting coordinates for the refinement were those read from a model fitted to a 2.8 Å mir map, whereas for insulin the coordinates were read by inspection from a 1.5 Å map, phased by the phase refinement method of Sayre (1971, 1974; Cutfield, Dodson, Dodson, Hodgkin, Isaacs, Sakabe & Sakabe, 1975). It now appears that if the lower resolution map is of sufficient quality, there is little to be gained by extending it to a higher resolution in order to improve the coordinates prior to refinement.

For a protein crystallographer, the computing requirements of the program are very modest. The program written by Dodson is flexible in its core storage requirements, and for the actinidin refinement (1820 atoms, 24000 data) required 35K words of store. It does need back up store, preferably on a disc, but could operate with magnetic tape files. The cost of the refinement, both in computer time and manpower, is its greatest attraction. The complete refinement of actinidin, from a set of coordinates read from a 2.8 Å map to a set of refined coordinates with 1.7 Å data took about 14 hrs of computer time on a DEC10 and was completed in only three months.

The Use of the Method

Isaacs and Agarwal (1978) and Baker and Dodson (1979) have recorded their experiences in using the method to refine insulin and actinidin respectively. Generally these experiences and those of Hardman with myoglobin and carbonic anhydrase (personal communication) are similar and it seems that differences in the size of the problem do not influence the nature of the refinement. The major difficulty with the method is the fact that the shifts calculated for geometrically related atoms, such as those forming a peptide unit, destroy the geometry. This behaviour is characteristic of protein refinements where atoms are allowed to move individually. Causes of this might be the large initial errors in the coordinates, the relative sparseness of the data with fewer than three observations for each variable parameter, and the neglect of atom-atom interactions in the normal matrix. This loss of geometry may be controlled using a program of the type written by Dodson, Isaacs and Rollet (1976) or Ten Eyck, Weaver and Mathews (1976) to correct the gross structural irregularities. Although this method of correcting the geometry every few cycles decreases the rate of convergence, there is no evidence so far to suggest that it adversely affects the accuracy of the final model. Test calculations performed by Stenkamp and Jensen (1976) with simulated protein data support this. A faster rate of convergence could be achieved by incorporating the geometrical constraints in the least-squares equations (Konnert 1976) but an advantage in separating the least-squares refinement and regularisation procedures is that large shifts on regularisation often indicate gross errors in the structure.

The experience of Isaacs and Agarwal (1978) and Baker and Dodson (1979) with high resolution data indicates that the gross errors in the model may be corrected by refinement with data to a resolution of 2 Å. The use of the weighting scheme proposed is important in placing most weight on the low angle terms for these early cycles. Baker and Dodson (1978) found that the initial seven cycles of coordinate refinement with 2 Å data on actinidin produced an average shift of 0.43 Å for main chain atoms. The remaining 21 cycles of coordinate refinement with the inclusion of data to 1.7 Å produced an average shift of 0.18 Å for the main chain atoms. Much of the manual labour required for a refinement is spent on the poorly defined regions of the structure and on the solvent structure. The well ordered solvent should be included in the model structure as soon as possible, but with regard to other solvent and to disordered structures it is wise to proceed with caution. Isaacs and Agarwal (1978) have discussed how incorrectly assigned solvent (water) molecules confused the interpretation of difference-Fourier densities of some side chains, particularly of glutamic acid and arginine residues. The solvent structure can be unravelled (Watenpaugh, Margulis, Sieker & Jensen, 1978) but to do so requires considerable effort.

Conclusion

Table 1 lists a number of structures which have been subject to refinement using the FFT least-square refinement programs. The speed of the new algorithm is evident – the complete refinement of actinidin, starting with coordinates from a 2.8 Å m.i.r. phased electron density map, required about 14 hours of computer time on a DEC 10 and was completed in only three months. This work also provides a good estimate of the radius of convergence of the method. The average shift in position for main chains was 0.45 Å and for side chains atoms 0.56 Å.

Acknowledgement

This tutorial presentation is based on material written by Dr. Neil Isaacs of St. Vincent's School of Medical Research, Melbourne, Australia. The author is greatly indebted to him for his contribution.

References

Adman, E.T., Sieker, L.C. and Jensen, L.H. (1975), Acta Cryst. A31, S34.

Agarwal, R.C. (1978), Acta Cryst. A34, 791-809.

Baker, E.N. and Dodson, E.J. (1979), J. Mol. Biol. in press.

Bode, E. and Schwager, P. (1975), J. Mol. Biol. 98, 693-717.

Chambers, J.L. and Stroud, R.M. (1977), Acta.Cryst. B33, 1824-1837.

Cooley, J.W. and Tukey, J.W. (1965), Math. Comput. 19, 297-301.

Cutfield, J.F., Dodson, E.J., Dodson, G.G., Hodgkin, D.C., Isaacs, N.W., Sakabe, K. and Sakabe, N. (1975), Acta Cryst. A31, S21.

Deisenhofer, J. and Steigemann, W. (1975), Acta Cryst. B31, 238-250.

Diamond, R. (1971), Acta Cryst. A27, 436-452.

Diamond, R. (1974), J. Mol. Biol. 82, 371-391.

Dodson, E.J., Isaacs, N.W. and Rollett, J.S. (1976), Acta Cryst. A32, 311-315.

Freer, S.T., Alden, R.A., Carter, C.W. and Kraut, J. (1975), J. Biol. Chem. 250, 46-54.

Hardman, K.D. (1978), Acta Cryst. A34, S65.

Huber, R., Kukla, D., Bode, W., Schwager, P., Bartels, K., Deisenhofer, J. and Steigemann, W. (1974), J. Mol. Biol. 89, 73-101.

Hull, S.E., Karlsson, R., Main, P., Woolfson, M.M. and Dodson, E.J. (1978), Nature, 275, 206-207.

International Tables for X-ray Crystallography (1974), Vol. IV, Birmingham: Kynoch Press.

Isaacs, N.W. and Agarwal, R.C. (1978), Acta Cryst. A34, 782-791.

Jack, A. and Levitt, M. (1978), Acta Cryst. A34, 931-935.
Konnert, J.H. (1976), Acta Cryst. A32, 614-637.
Moews, P.C. and Kretsinger, R.H. (1975), J. Mol. Biol. 91, 201-228.
Sayre, D. (1951), Acta Cryst. 4, 362-367.
Sayre, D. (1972), Acta Cryst. A28, 210-212.
Sayre, D. (1974), Acta Cryst. A30, 180-184.
Schmidt, Jr., W.C., Girling, R. L. and Amma, E.L. (1977), Acta Cryst. B33, 3618-3620.
Stenkamp, R.E. and Jensen, L.H. (1976), Acta Cryst. A32, 255-258.
Sussman, J.L., Holbrook, S.R., Chruch, G.M. and Kim, S.H. (1977), Acta Cryst. A33, 800-804.
Takano, T. (1977), J. Mol. Biol. 110, 537-584.
Ten Eyck, L.F. (1973), Acta Cryst. A29, 183-191.
Ten Eyck, L.F. (1977), Acta Cryst. A33, 486-492.
Ten Eyck, L.F., Weaver, L.H. and Mathews, B.W. (1976), Acta Cryst. A32, 349-350.
Watenpaugh, K.D., Sieker, L.C., Herriott, J.R. and Jensen, L.H. (1973), Acta Cryst., B29, 943-956.
Watenpaugh, K.D., Margulis, T.N., Sieker, L.C. and Jensen, L.H. (1978), J. Mol. Biol. 122, 175-190.
Winograd, S. (1978), Mathematics of Computation, 32, 175-199.

TABLE 1

Some examples of structures refined by FFT least-squares

Molecule	No. atoms	No. data	d_{min}(Å)	Space Gp.	R factors Ini- tial	R factors Fi- nal	No. Cycles XYZ	No. Cycles B	CPU/ cycle*	Computer	Refe- rence
Insulin	1077	11890	1.5	R3	.28	.11	43	24	3 min	IBM 370/168	a
"	"	"	"	"	-	-	-	-	12 min	DEC 10	a
6-acetyl-dolatriol	26	1041	1.0	R3	.50	.10	13	2	20 s.	IBM 370/168	b
Beavuricin Ba salt	370	6667	1.2	$P2_1$	.21	.12	14	10	38 s.	"	b
Metmyoglobin	∿1400	∿9000	2.0	"	.27	.16	5	3	68 s.	"	c
Carbonic anhydrase	2050	9451	2.0	"	.42	.18	9	5	-	"	d
Actinidin	1821	23390	1.7	$P2_12_12_1$	.43	.17	28	14	20 m.	DEC 10	e
"	"	"	"	"	-	-	-	-	10 m.	CYBER 74-16	e
Gramicidin S	84	4902	1.0	$P3_121$	-	.19	-	-	-	DEC 10	f

* Cpu times for the DEC 10 include a charge for operating costs

a. Isaacs and Agarwal (1978)
b. Agarwal (1978)
c. Hardman (1978)
d. Hardman, K.D. personal communication.
e. Baker and Dodson (1979)
f. Hull, Karlsson, Main, Woolfson and Dodson (1978).

PHASE EVALUATION AND SOME ASPECTS OF THE FOURIER REFINEMENT OF MACROMOLECULES

M. Vijayan

Molecular Biophysics Unit, Indian Institute of Science
Bangalore 560 012, India

SUMMARY

The statistical method for the determination of protein phases from isomorphous and anomalous differences has been described. Methods for estimating r.m.s. errors and the suggested modifications to the classical Blow and Crick formulation have been reviewed. The Fourier methods for the refinement of protein structures have been outlined. A theoretical analysis, with special reference to protein structures, has been presented. The theory provides a rationalisation for the use of general Fourier syntheses with $(mF_o - nF_c)\exp(i\alpha_c)$ as coefficients; it also leads to the determination of optimum values of m, n and other related parameters. A method for the empirical determination of the parameters has been suggested. The treatment of "inner reflections" affected by solvent, the reliability of refined coordinates and the checking procedures employed during the course of the refinement have also been discussed.

INTRODUCTION

Despite the striking advances made in recent years in the development of novel methods and procedures for the x-ray study of macromolecules, isomorphous replacement, which is often used in conjunction with anomalous scattering data, remains the most important and almost indispensible method of phase determination in macromolecular crystallography. This contribution, therefore, starts with a description of the classical phase evaluation procedures using isomorphous and anomalous differences; some suggested improvements to the classical formulation are also touched upon.

Until recently, macromolecular crystallography has been concerned primarily with the determination of the gross three-dimensional structure of macromolecules from isomorphously phased electron-density maps.

Several successful attempts have, however, been made during the last few years to refine macromolecular structures to the maximum accuracy permitted by the total available data from the native crystals. The excellent set of articles on the high resolution refinement of protein structures in the Proceedings of the 1975 International Summer School on Crystallographic Computing Techniques held at Prague provides a nearly comprehensive account of the state of the work in this area as it then existed. There have been several subsequent developments some of which are discussed in other contributions in this Winter School. I shall be concerned, in the latter part of this contribution, with the Fourier methods of macromolecular structure refinement. No attempt will, however, be made to give a comprehensive account of the application of these methods. Instead, I shall endeavour to outline the methods employed, present a theoretical analysis, highlight some special problems and suggest some improvements in the existing procedures.

The crystallographic techniques developed for the structure analysis of proteins can be, and have been, used for the study of other macromolecules as well. However, as most of the macromolecular crystallographic studies have till now been concerned with proteins, the term "protein" will be used here, for the sake of convenience, to refer to macromolecules in general.

PHASE EVALUATION

Isomorphous replacement and anomalous dispersion methods

The preparation of isomorphous protein heavy atom derivatives involves the attachment of "heavy" atoms like mercury, lead or uranium or chemical groups containing them to protein crystals in a coherent manner without changing the conformation of the molecules and their crystal packing. Thus, ideally, the structures of a protein crystal and a derivative crystal should be identical as far as the ordered regions are concerned except for the presence of heavy atoms or heavy atom containing groups in the latter. Under such conditions of ideal isomorphism, and neglecting experimental errors, the relation between the structure factors of any given acentric reflection from the native crystal and the derivative crystal can be represented as in Figure 1. In the diagram, the magnitudes of the structure factor of the heavy atom derivative F_{PH}, the structure factor of the protein F_P and the heavy atom contribution F_H are denoted by F_{PH}, F_P and F_H respectively. The respective phase angles are α_{PH}, α_P and α_H. Of these, F_{PH} and F_P can be obtained directly from experimental data. F_H can be calculated from the refined heavy atom parameters. The phase angle of the protein

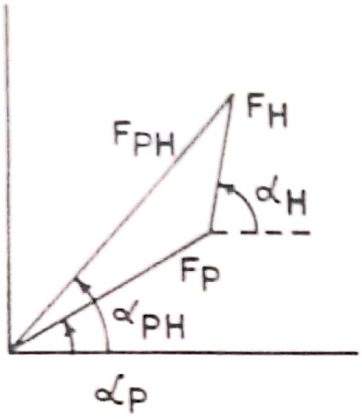


Figure 1

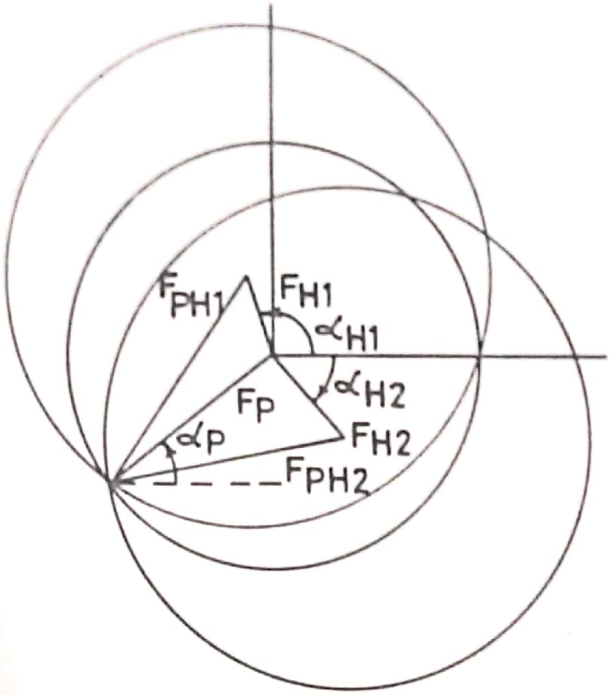


Figure 2

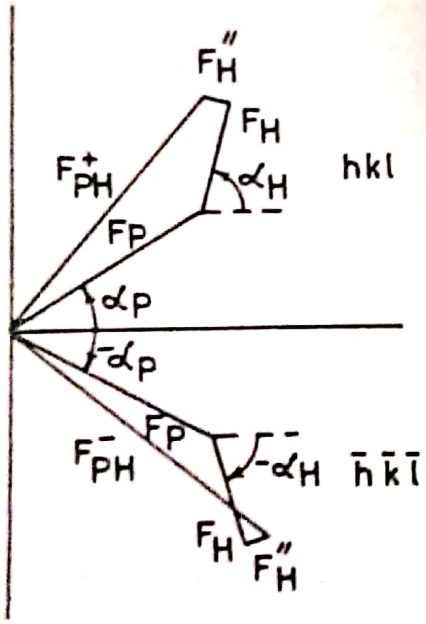


Figure 3

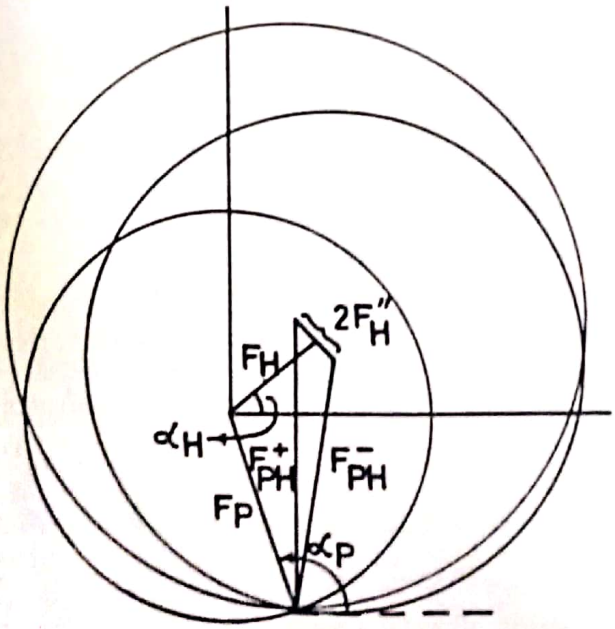


Figure 4

structure factor is then given by

$$\alpha_P = \alpha_H \pm \phi \qquad (1)$$

where $\cos(\pi-\phi) = (F_{PH}^2 - F_P^2 - F_H^2)/2F_PF_H$. Thus there are two possible values for α_P placed symmetrically about α_H. This ambiguity can be resolved if data from two independent derivatives are available. Two equations like (1) would then be available.

$$\alpha_P = \alpha_{H1} \pm \phi_1$$

and

$$\alpha_P = \alpha_{H2} \pm \phi_2 \quad ,$$

where subscripts 1 and 2 refer to derivatives 1 and 2 respectively. There are thus two possible sets of values. That value which is common to both the sets corresponds to the correct protein phase angle. The situation can be demonstrated graphically with the aid of the so called Harker construction (Harker, 1956) shown in Figure 2. A circle is drawn with F_P as radius and the origin of the vector diagram as the centre. Two more circles are drawn with F_{PH1} and F_{PH2} as radii and the ends of the vectors $\bar{F}_{H1}$ and $\bar{F}_{H2}$ as the centres. Both the circles intersect the F_P circle at two points each. One of the points of intersection is common. That point defines the phase angle of the structure factor from the native crystal. Thus protein phase angles can be determined if a minimum of two independent heavy atom derivatives are available.

The dispersion correction terms (International Tables for X-ray Crystallography, 1962) for atoms with high atomic numbers are appreciable and hence the heavy atoms in protein derivatives are usually anomalous scatterers. Assuming that the heavy atoms in the derivatives are the only anomalous scatterers and that all the heavy atoms in any given derivative are of the same type, the relation between the structure factor of a reflection hkl from a derivative and that of its Friedel partner $\bar{h}\,\bar{k}\,\bar{l}$ can be represented as in Figure 3. The magnitudes of the two structure factors are denoted by F_{PH}^+ and F_{PH}^- respectively. $\bar{F}_H$ is the real part of the heavy atom contribution including that due to the real part of the dispersion correction and $\bar{F}_H''$ is the imaginary component of the heavy atom contribution. It is readily seen that $\bar{F}_{PH}^+$ and $\bar{F}_{PH}^-$ could be formally considered as the structure factors of any given reflection arising from two independent derivatives. The Harker diagram can then be constructed as shown in Figure 4. Therefore, in principle, protein phase angles can be determined from a single derivative when anomalous scattering effects are also made use of.

It is interesting to note that, for any given derivative, the information obtained from isomorphous differences, $F_{PH}-F_P$, and that obtained from anomalous differences, $F^+_{PH} - F^-_{PH}$, are complementary. The isomorphous difference for any given reflection is maximum if $\overline{F_P}$ and $\overline{F_H}$ are parallel or antiparallel. The anomalous difference, then, is zero if all the anomalous scatterers are of the same type, and the native phase angle is determined uniquely on the basis of the isomorphous difference. In general, the isomorphous difference decreases and the anomalous difference increases as the inclination between $\overline{F_P}$ and $\overline{F_H}$ increases. Assuming F_H to be small compared to F_P, the isomorphous difference tends to be small and the anomalous difference tends to have its maximum possible value when $\overline{F_P}$ and $\overline{F_H}$ are perpendicular to each other. The anomalous difference then has a predominant influence in determining the phase angle.

Blow and Crick formulation

In a real situation, conditions are far from ideal on account of several factors the chief among them being imperfect isomorphism, errors in the estimation of heavy atom parameters and the experimental errors in the measurement of intensity from the native and the derivative crystals. Consequently it is desirable to use as many derivatives as are available for phase determination. All the circles would not then intersect at a single point in the Harker diagram; instead there would be a distribution of intersections. Thus what one obtains is not a unique phase angle, but a probability distribution for the phase angle.

The statistical procedure employed so far for deriving phase angles using multiple isomorphous replacement (MIR) is based on the classical treatment by Blow and Crick (1959). In their treatment, Blow and Crick assume, for mathematical convenience, that all errors could be considered as residing in the magnitude of the derivative structure factor alone. They make a further assumption that those errors could be described by a Gaussian distribution. With these simplifying assumptions, the statistical procedure for phase determination can be readily worked out as follows.

Figure 5 shows the vector diagram for a reflection from a particular derivative with an arbitrary value α for the protein phase angle. Referring to the Figure, we have for the derivative

$$D_{Hi}(\alpha) = \left\{F_P^2 + F_{Hi}^2 + 2F_P F_{Hi} \cos(\alpha_{Hi}-\alpha)\right\}^{1/2} \qquad (2)$$

If α corresponds to the true protein phase angle α_P, then $D_{Hi}(\alpha)$ coincides with F_{PHi}. The amount by which $D_{Hi}(\alpha)$ differs from F_{PHi}, namely,

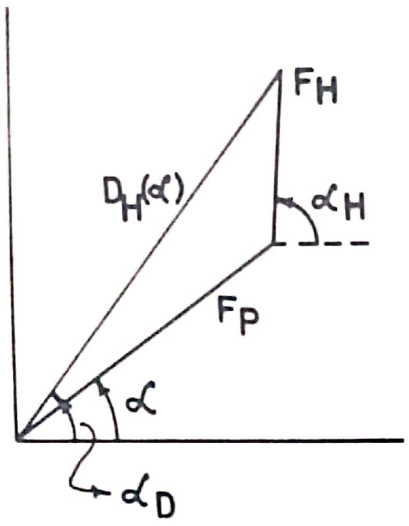


Figure 5

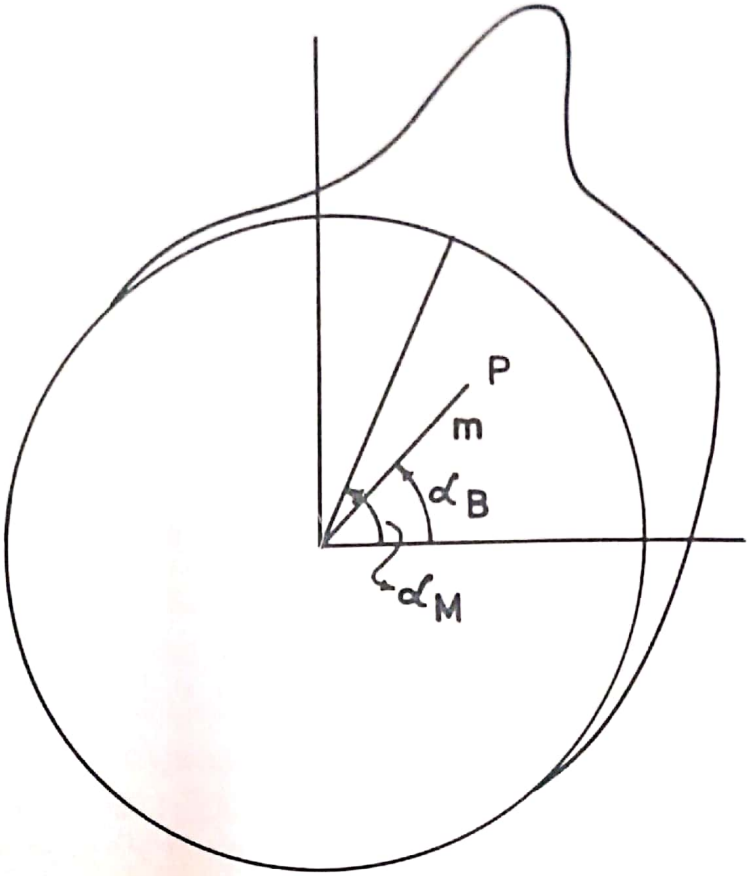


Figure 6

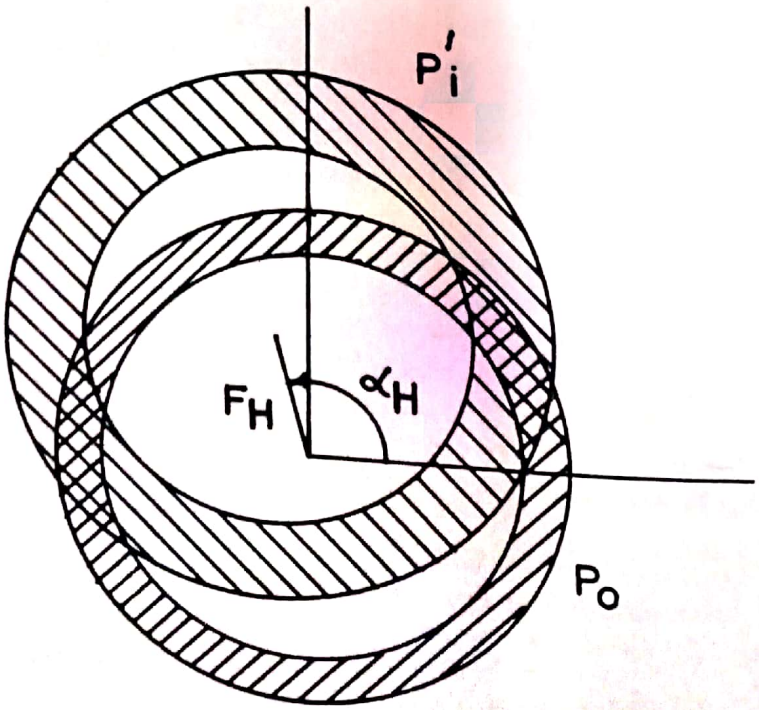


Figure 7

$$\xi_{Hi}(\alpha) = F_{PHi} - D_{Hi}(\alpha) \tag{3}$$

is a measure of the incorrectness of the phase angle α. This quantity is called the lack of closure. Then the probability for α being the correct protein phase angle is defined as

$$P_i(\alpha) = N_i \exp\left\{-\xi^2_{Hi}(\alpha)/2E_i^2\right\} \tag{4}$$

where N_i is the normalisation constant and E_i is an estimate of the r.m.s. error. When a number of heavy atom derivatives are available, the total probability of a phase angle α being correct would be

$$P(\alpha) = \pi P_i(\alpha) = N \exp\left\{-\sum_i (\xi^2_{Hi}(\alpha)/2E_i^2)\right\} \tag{5}$$

where the summation is over all the derivatives.

When $P(\alpha)$ for any particular reflection is plotted around a circle of unit radius, as shown in Figure 6, the phase corresponding to the highest peak in the probability distribution would give the most probable protein phase α_M for the reflection. Then the Fourier synthesis with

$$F_P \exp(i\alpha_M)$$

as coefficients would give the most probable electron-density distribution in the protein. A different way of using the probability distribution has been described by Blow and Crick. In Figure 6, the centroid of the probability distribution is at point P. The polar coordinates of P are m and α_B where m, a fractional positive number with a maximum value of unity, and α_B are referred to as the "Figure of merit" and the "best phase" respectively. A Fourier synthesis with

$$mF_P \exp(i\alpha_B)$$

as coefficients is called the "best Fourier". Defined in this manner, the best Fourier would give the electron-density distribution with the lowest r.m.s. error. The best Fourier synthesis rather than the most probable Fourier synthesis is usually employed in the structure analysis of proteins. In practice, the figure of merit and the best phase are calculated using the expressions

$$m \cos \alpha_B = \sum_i P(\alpha_i) \cos \alpha_i / \sum_i P(\alpha_i)$$

and

$$m \sin \alpha_B = \sum_i P(\alpha_i) \sin \alpha_i / \sum_i P(\alpha_i) \tag{6}$$

where $P(\alpha_i)$ are calculated, say, at 5° intervals (Dickerson et al., 1961). The figure of merit gives an estimate of precision of the calculated phase angle and it is statistically interpreted as the cosine of the expected error in the calculated phase angle. Obviously, it has a high value when α_M and α_B are close to each other and a low value when they are far apart.

In the presence of anomalous scattering data, when F^+_{PH} and F^-_{PH} are treated as arising from two independent derivatives, the effect of anomalous differences on phase determination would only be marginal as, for any given reflection, the difference between F^+_{PH} and F^-_H is usually small. North (1965) has, however, pointed out that the error in the anomalous difference for a given reflection would normally be much smaller than that in the corresponding isomorphous difference. First the former is obviously free from the effects of non-isomorphism. Secondly, as F^+_{PH} and F^-_{PH} are measured from the same crystal, both these quantities are expected to have the same systematic errors. These errors are eliminated in the difference between the two quantities. Therefore, different estimates of the root mean square error E should be used for isomorphous and anomalous differences. Then, for any given derivative, the new expression for the probability distribution of the protein phase angle in the presence of anomalous scattering data would be

$$P_i(\alpha) = N_i \exp\left\{-\xi^2_{Hi}(\alpha)/2E_i^2\right\} \exp\left\{-(\Delta H_i - \Delta H_{ical})^2/2E_i'^2\right\} \qquad (7)$$

where

$$\Delta H_i = F^+_{PHi} - F^-_{PH}, \quad \Delta H_{ical} = 2F''_{Hi} \sin(\alpha_D - \alpha_H)$$

is the corresponding value of the anomalous difference calculated for the phase angle α and E'_i is the r.m.s. error in anomalous differences. In the derivation of (7), F_{PHi} is taken as the average value of F^+_{PHi} and F^-_{PHi}, and F_{Hi} is approximated to be the magnitude of the real part of the heavy atom contribution including the real component of the dispersion correction.

Estimation of E and E'

E and E' are two important parameters that have to be estimated for computing phase angles. Blow and Crick suggest the use of centric reflections, when present, for evaluating E using the expression

$$E^2 = \sum_n (|F_{PH} \pm F_P| - F_H)^2/n \qquad (8)$$

The values of E (one for each derivative) thus obtained can be used for acentric reflections as well. When preliminary estimates of phase angles are available and when the number of derivatives is large, E can be calculated as the r.m.s. lack of closure corresponding to α_B (Kartha, 1976). The r.m.s. error in anomalous differences can also be evaluated by similar methods. In general, the value of E' for any given derivative is about a third of the corresponding value of E.

The values of E and E' for each derivative can also be evaluated by a different method (Adams, 1968) as outlined below using the following well known relations (Ramachandran and Raman, 1956).

$$\cos \alpha = -(F_P^2 - F_{PH}^2 - F_H^2)/2F_{PH}F_H \tag{9}$$

and

$$\sin \alpha = (F_{PH}^{+2} - F_{PH}^{-2})/4F_{PH}F_H'' \tag{10}$$

where $\alpha = \alpha_{PH} - \alpha_H$. We obtain what may be called α_{iso} if the magnitude of α is determined from (9) and the quadrant from (10). Similarly, we obtain α_{ano} if the magnitude of α is determined from (10) and quadrant from (9). The difference between α_{iso} and α_{ano} is a measure of the errors present in the data. From (9) we have

$$F_P^2 = F_{PH}^2 + F_H^2 - 2F_{PH}F_H \cos \alpha \tag{11}$$

Using α_{ano} in (11) we obtain what may be considered as the calculated value of F_P (F_{Pcal}). Assuming that all errors lie in F_P, the values of E can be calculated using the expression

$$E^2 = \sum_n |F_P - F_{Pcal}|^2/n \tag{12}$$

Now assuming F_{PH} to be equal to $(F_{PH}^+ + F_{PH}^-)/2$, we have from (10)

$$F_{PH}^+ - F_{PH}^- = 2F_H'' \sin \alpha \tag{13}$$

The value obtained by using α_{iso} in (13) may be considered as the calculated value of the anomalous difference (ΔH_{cal}). The values of E' can then be evaluated using the expression

$$E'^2 = \sum_n |\Delta H - \Delta H_{cal}|^2/n \tag{14}$$

The values of E can also be evaluated from acentric data by making use of the values of F_H estimated by combining isomorphous and anomalous differences. Kartha and Parthasarathy (1965) and Mathews (1966) have given approximate formulae for estimating F_H; the exact formula was subsequently derived by Singh and Ramaseshan (1966). The estimate of F_H, however, is not unambiguous. For every reflection.

there are two possible estimates, an upper estimate (F_{HUE}) and a lower estimate (F_{HLE}). Under normal conditions, F_{HLE} would correspond to the correct estimate for most reflections (Dodson and Vijayan, 1971). Now assuming that all errors lie in F_H, the values of E can be estimated from the expression

$$E^2 = \sum_n |F_{HLE} - F_H|^2/n \qquad (15)$$

The root mean square error E (and also E' when anomalous differences are used) is an important parameter in phase determination. For a given derivative the sharpness of the peak(s) in the probability distribution obviously depends on the choice of E. When several derivatives are used, an overall decrease in the values of E from their correct values leads to artificially sharper peaks, the movement of α_B towards α_M and deceptively high figures of merit. Opposite effects result from an increase in the values of E. It is also important to see that the estimated E in each derivative is a correct measure of the r.m.s error for that particular derivative to ensure the correct relative contribution from the derivative to the overall phase probability distribution.

Suggested modifications to Blow and Crick formulation

Ashida (1976) has discussed some modifications to the Blow and Crick procedure while retaining its essential characteristics in form as well as in content. In the first modification, originally proposed by Cullis et al. (1961), discussed by him, all the E_i's are assumed to be the same and the lack of closure error ξ_{Hi} for the i^{th} derivative is measured as the distance from the mean of all intersections between phase circles to the point of intersection of the phase circle of that derivative with the phase circle of the native protein. In the second modification discussed by him, individual values of E_i for different derivatives are retained; but the lack of closure is measured from the weighted mean of all intersections.

Another modification, again within the framework of the Blow and Crick formulation, was earlier proposed by Hendrickson and Lattman (1970). The Blow and Crick procedure is based on the relation

$$|\overline{F_P} + \overline{F_{Hi}}| = F_{PHi} + \xi_{Hi} \qquad (16)$$

where ξ_{Hi} is the "lumped" error, assumed to be Gaussian, in F_{PHi}. Hendrickson and Lattman instead use the relation

$$|\overline{F_P} + \overline{F_{Hi}}|^2 = F^2_{PHi} + \xi''_{Hi} \qquad (17)$$

where ξ''_{Hi} is the lumped error, again assumed to be Gaussian, in F_{PHi}. The corresponding r.m.s. error, E''_i, can be evaluated using methods similar to those employed for evaluating E_i. Hendrickson and Lattman point out that whereas the values of E have been shown to be only slightly dependent on the measured intensities, the values of E'' would necessarily be functions of structure factor amplitudes. The real advantage in using the modified procedure lies in the fact that the exponent in the probability expression can then be expressed as a linear combination of five terms in the following manner

$$-\xi''^2_{Hi}(\alpha)/2E''^2_i = K_i + A_i \cos\alpha + B_i \sin\alpha + C_i \cos 2\alpha + D_i \sin 2\alpha \qquad (18)$$

where K_i, A_i, B_i, C_i and D_i are constants dependant on F_P, $\overline{F_{Hi}}$, F_{PHi} and E''_i. The complete probability distribution of any reflection can thus be expressed in terms of five constants. Similar expressions have been derived for phase information from anomalous scattering, tangent formula, partial structure and non-crystallographic symmetry. The phase information from all sources can then be combined by simply taking the total value of each constant. Thus, the total probability of the phase angle to be α is given by

$$P(\alpha) = \pi P_s(\alpha) = N' \exp\Big(\sum_s K_s + \sum_s A_s \cos\alpha + \sum_s B_s \sin\alpha + \sum_s C_s \cos 2\alpha + \sum_s D_s \sin 2\alpha\Big) \qquad (19)$$

where K_s, A_s, etc. are the constants appropriate to the s^{th} source and N' is the normalisation constant.

When the total probability of the phase angle being α is represented as

$$P(\alpha) = \pi P_i(\alpha) \ ,$$

the individual probabilities obtained from different derivatives are assumed to be independent and hence are multiplied to get the total probability. This follows from the assumption that all errors reside in F_{PH}. However, in fact, F_P, F_H and F_{PH} should all be considered in error. If the probability related to errors in F_P is denoted by P_o, Einstein (1977) points out that each P_i involves the term P_o. Therefore, different P_i are not independent and hence should not be multiplied. The effect of multiplying them is to give too high a weight to the observed F_P. This effect is sought to be eliminated by Raiz and Andreeva (1970) and Einstein (1977) by the explicit consideration of errors in F_P as well. Following Einstein, the basic principle of their procedure can be illustrated diagramatically as shown in Figure 7. The Harker diagram for any given derivative no longer consists of two intersecting circles. The parent circle is replaced by the probability distribution P_o which is

indicated as a shaded annular region between circles of equi-probability contours. The probability distribution P_i', related to errors in F_{PH} and F_H, is also shown in a similar manner. The joint probability distribution $P_o P_i'$ can now be used instead of P_i. $P_o P_i'$ is multiplied by a distribution of the type P_i' for each additional derivative. Mathematical formulae for describing such joint probability distributions have been derived by Einstein. A procedure for including anomalous scattering data, within the framework of his formulation, has also been described.

Perhaps the most comprehensive set of modifications to the Blow and Crick formulation is that suggested by Green (1979) although the treatment is limited to the case of a single derivative. Errors arising from imperfect isomorphism, errors in the heavy atom positions and those associated with $\overline{F_{PH}}$ and $\overline{F_P}$ are separately considered in this treatment. Probability formulae for situations where errors of each type or all types are present have been derived. Methods have also been suggested for estimating the r.m.s. value of each type of error. Perhaps the most interesting result of Green's analysis is the treatment of imperfect isomorphism. The analysis shows that the reliability of phase estimation decreases with increasing $\sin\theta/\lambda$ when isomorphism is assumed to be imperfect. This is indeed what one would expect on physical grounds as the effects of departures from strict isomorphism are likely to be important at high resolution where the d spacings tend to be comparable to the magnitudes of such departures.

FOURIER REFINEMENT OF PROTEIN STRUCTURES

Methods

Most, but by no means all, of the refinements performed to date on protein structures have been carried out at resolutions of 2 Å or better on models obtained from MIR phased maps at resolutions in the range of 2 to 3 Å. The Fourier methods so far employed in these refinements can be broadly classified into two categories, namely, that developed and used mainly by the Munich group and that used by most others.

The Munich group has refined several protein structures (Huber et al. 1974; Deisenhoffer and Steigemann, 1975; Epp et al. 1975; Bode and Schwager, 1975) using the procedure outlined below. In their method, a synthesis with coefficients

$$(nF_o - (n-1)F_c)\exp(i\alpha_c) \quad , \qquad (20)$$

where F_o is the magnitude of the observed structure factor, and F_c and α_c are the magnitude and the phase angle of the structure factor calculated

from the current set of coordinates (starting set for the first cycle), is computed in each cycle of refinement. The model is then fitted to the map using the well-known real space refinment procedure developed by Diamond (1971). The resulting coordinates are then used as input parameters for the next cycle of refinement which again involves the calculation of a Fourier map with (20) as coefficients and subsequent real space refinement. The automatic procedure is interrupted, when appropriate, to correct gross errors and to locate solvent molecules using conventional difference Fourier (ΔF) maps with

$$(F_o - F_c)\exp(i\alpha_c) \tag{21}$$

as coefficients. A new cycle of refinement then starts with the corrected coordinates. The cyclic procedure is stopped when convergence in terms of R factor and parameter shifts is reached, and the ΔF map is flat.

The method employed by most others (Watenpaugh et al. 1973; Freer et al. 1975; Moews and Kretsinger, 1975; Chambers and Stroud, 1977; Stenkamp et al. 1978; Blake et al. 1978; Dodson et al. 1979) is essentially the same as that used in the refinement of small molecular structures. The ΔF synthesis is cyclically used for correcting positional and thermal parameters. Shifts in positional parameters are calculated using the relation

$$\delta x_i = -\,\text{gradient/curvature} = -\frac{\partial \Delta\rho / \partial x_i}{\partial^2 \rho / \partial x_i^2} \tag{22}$$

where δx_i is the shift in the parameter x_i. The gradients in the ΔF map at the assumed atomic positions are estimated by interpolation of different densities at the surrounding grid points. The approximate curvatures of different atoms are estimated by one empirical method or the other from an electron-density (F_o) map. A variety of empirical formulae have been used for estimating shifts in thermal parameters; all of them naturally seek to increase the B value if the difference density at the assumed atomic position is negative and decrease the B value if it is positive. As the resolution of the data is limited, the shifted parameters do not, in general, lead to acceptable molecular geometries. Therefore, the protein molecule is regularised, in between cycles, using one of the available automatic methods (Diamond, 1966; Hermans and McQueen, 1974; Dodson et al. 1976) to restore molecular dimensions to within acceptable limits.

A theoretical analysis

It is convenient to discuss some of the problems associated with the Fourier refinement of protein structures in terms of the following theoretical analysis. If a protein structure contains a total of N atoms (including those in solvent molecules), at any given stage in the course of refinement, the inaccurate positions $\bar{r}'_{Pj}$ of P atoms are known whereas the positions $\bar{r}_{Qj}$ of the remaining Q atoms are unknown. The true positions of the P known atoms may be denoted by $\bar{r}_{Pj}$. The calculated structure factors corresponding to $\bar{r}_{Pj}$ and $\bar{r}'_{Pj}$ may be denoted by $F_P \exp(i\alpha_P)$ and $F'_P \exp(i\alpha'_P)$, and the magnitude of the observed structure factor by F_N. Obviously, F_N, F'_P and α'_P correspond to the conventional F_o, F_c and α_c respectively. We also define

$$S_N^2 = \sum_{j=1}^{N} f_{Nj}^2$$

$$S_P^2 = \sum_{j=1}^{P} f_{Pj}^2 \qquad (23)$$

and $$S_Q^2 = \sum_{j=1}^{Q} f_{Qj}^2$$

where f_{Xj} is the scattering factor for atom Xj. f_{Xj} is assumed to have been corrected for the temperature factor. Obviously, $N = P + Q$ and $S_N^2 = S_P^2 + S_Q^2$. Then, following the methods developed by Ramachandran and Srinivasan (1970), it can be shown (Vijayan, 1980) that a general Fourier synthesis with coefficients

$$(mF_N - nF'_P) \exp(i\alpha'_P) \qquad (24)$$

has the following peak positions and strengths

$$\bar{r}'_{Pj} \qquad \frac{mS_N - nS_P}{S_P} f_{Pj} \qquad (25)$$

$$\bar{r}'_{Pj} + \bar{r}'_{Pk} - \bar{r}'_{Pl} \qquad -\frac{mS_N}{2S_P^3} f_{Pj} f_{Pk} f_{Pl} \qquad (26)$$

$(j \neq l)$

$$\bar{r}'_{Pj} + \bar{r}_{Qk} - \bar{r}_{Ql} \quad (k \neq 1) \qquad \frac{m}{2S_N S_P} f_{Pj} f_{Qk} f_{Ql} \tag{27}$$

$$\bar{r}'_{Pj} + \bar{r}_{Pk} - \bar{r}_{Ql} \qquad \frac{m}{2S_N S_P} f_{Pj} f_{Pk} f_{Ql} \tag{28}$$

$$\bar{r}_{Pj} + \bar{r}'_{Pk} - \bar{r}_{Pl} \quad (j \neq 1) \qquad \frac{m}{2S_N S_P} f_{Pj} f_{Pk} f_{Pl} \tag{29}$$

$$\bar{r}_{Qj} + \bar{r}'_{Pk} - \bar{r}_{Pl} \qquad \frac{m}{2S_N S_P} f_{Qj} f_{Pk} f_{Pl} \tag{30}$$

The above distribution of peaks can be considered as consisting of shifted vector sets centred around different true or assumed atomic positions, some with origin peaks and some without. Terms (27) and (28), and terms (26), (29) and (30) when $k \neq 1$ give rise to general background. (26), (29) and (30) give rise to peaks at assumed and true atomic positions when $k=1$ and when the errors in the positions of the known P atoms are small and random. Under such conditions and assuming N and P to be large, which is true in the case of proteins, the peaks at atmoc positions will have the following strengths.

$$\bar{r}_{Pj} \qquad \frac{mS_N - 2nS_P}{2S_P} f_{Pj} \tag{31}$$

$$\bar{r}_{Pj} + \langle (\bar{r}'_{Pk} - \bar{r}_{Pk}) \rangle \quad (J \neq k) \qquad \frac{mS_P}{2S_N} f_{Pj} \tag{32}$$

$$\bar{r}_{Qj} + \langle (\bar{r}'_{Pk} - \bar{r}_{Pk}) \rangle \qquad \frac{mS_P}{2S_N} f_{Qj} \tag{33}$$

The conventional ΔF map results when $m=n=1$; likewise conventional F_o map results when $m=1$ and $n=0$. It can be readily shown that the known properties of these maps (Luzzati, 1953; Dodson and Vijayan, 1971) follow from the above three expressions.

Parametrisation

As indicated earlier, two types of syntheses, one with (20) as coefficients and the other the ΔF synthesis, have been used in the Fourier refinement of protein structures. When using the former, one seeks to obtain a true representation of the electron-density distribution in the unit cell with peak strengths of f_{Pj} and f_{Qj} at $\bar{r}_{Pj}$ and $\bar{r}_{Qj}$ respectively

and zero density at $\bar{r}'_{Pj}$. This is readily achieved in a synthesis with

$$(mF_o - nF_c)\exp(i\alpha_c) \tag{34}$$

as coefficients when

$$m = 2S_N/S_P \quad \text{and} \quad n = S_N^2/S_P^2 \tag{35}$$

The above synthesis is identical, except for a scale factor, to a synthesis with

$$(kF_o - F_c)\exp(i\alpha_c) \tag{36}$$

as coefficients. The best results are obtained from this synthesis when

$$k = m/n = 2S_P/S_N \ . \tag{37}$$

Thus the theory outlined earlier provides a rationalisation for syntheses with a linear combination of F_o and F_c as the magnitude of the coefficients; it also leads to the determination of the optimum values of the parameters (m, n or k) to be used. The parametrisation can be made still more effective by using empirically evaluated values of S_N and S_P as will be shown later.

The ΔF synthesis is most effective when P is nearly equal to N. The peak strengths at $\bar{r}'_{Pj}$, $\bar{r}_{Pj}$ and $\bar{r}_{Qj}$ are then expected to be $-f_{Pj}/2$, $f_{Pj}/2$ and $f_{Qj}/2$ respectively when all reflections are acentric. The peaks with equal magnitudes and opposite signs at $\bar{r}'_{Pj}$ and $\bar{r}_{Pj}$ combine to give a density gradient which is made use of (as in (22)) to evaluate shifts in positional parameters. In practice P and N differ significantly except in the very final stages of refinement and, consequently, peaks at $\bar{r}'_{Pj}$ and $\bar{r}_{Pj}$ will have reduced and unequal strengths in the ΔF synthesis. However, peak strengths at the atomic positions expected in a ΔF map when $P \sim N$ can be reproduced in a map with

$$(m'F_o - n'F_c)\exp(i\alpha_c) \tag{38}$$

as coefficients if the parameters are chosen as

$$m' = S_N/S_P \quad \text{and} \quad n' = \tfrac{1}{2} + S_N^2/S_P^2 \tag{39}$$

Here again the synthesis can be made more effective by the use of empirically evaluated S_P and S_N (see later).

Empirical values of S_P and S_N

Unlike in the structure of most small molecules, the definition of the structure in protein crystals vary substantially from one region of the asymmetric unit to another. In general, most of the main chain atoms and the atoms belonging to internal side chains are well defined with low "temperature factors". The atoms belonging to surface residues and solvent molecules are usually poorly defined. They are often associated with high temperature factors arising from large thermal vibration amplitudes as well as static disorder corresponding to different structural or conformational possibilities. Consequently these atoms are associated with weak and diffuse electron-densities. The positions of the well-defined atoms are most often located in the early stages of refinement and attempts are then made to locate poorly defined atoms. Therefore, at any given stage in the course of the refinement, the scattering power of the P known atoms are likely to be different from those of the remaining Q atoms. Thus, S_N^2 and S_P^2 would not correspond to the true scattering power of the complete structure and the known part of the structure respectively unless the form factor of each atom is properly corrected for temperature factor. Some estimate of the temperature factors of the known atoms may exist; no such estimate for those of the unknown atoms will be available. Therefore, it is advisable to evaluate S_N and S_P empirically. Assuming Wilsonian distribution of intensities, this can be done by replacing S_N^2 and S_P^2 by $\langle F_N^2 \rangle$ and $\langle F_P^2 \rangle$ respectively.

Yet another factor that need to be considered is the differential contribution of the P and the Q atoms to intensities at different Bragg angles. If the Q atoms belong mostly to surface residues and solvent molecules, which is usually the case, their contribution, though high at low resolution, is likely to decrease rapidly with increasing Bragg angle. The differential fall off of the scattering powers of the known and the unknown parts of the structure can be taken care of by dividing the reciprocal space into a convenient number of spherical shells with increasing radii ($4 \sin^2 \theta/\lambda^2$) and then evaluating S_N^2 ($=\langle F_N^2 \rangle$) and S_P^2 ($=\langle F_P^2 \rangle$) in each shell separately. One can then obtain a curve for each parameter (m, n, k, m' or n') as a function of Bragg angle. For each reflection, the values of the parameters for the corresponding Bragg angle can be used to construct the appropriate Fourier coefficient.

Treatment of inner reflections

Problems concerned with the treatment of reflections with very low Bragg angles (say, in the 6 Å sphere) have been discussed in most of the

reports on the refinement of protein structures. These are the reflections most seriously affected by the presence of disordered solvent regions in the crystal. Many of these solvent molecules, if not most of them, are not, or cannot be, included in the structure factor calculations. Their scattering power at high angles are likely to be small on account of thermal or static disorder. Their contribution to reflections at very low angles would, however, be substantial. Therefore, the effect of the unknown part of the structure, made up substantially of solvent molecules, is most pronounced for the inner reflections. Thus, large discrepancies between the observed and the calculated structure factors are expected for these reflections. The observed discrepancies, however, appear to be systematic. In almost all cases, the calculated structure factors are reported to be much greater than the observed ones.

A satisfactory, though qualitative, explanation of this phenomenon, based on Babinet's principle, has been given by Moews and Kretsinger (1975). If the scattering matter is uniformly distributed in the unit cell, the vector sum of the scattering amplitude from one region of the unit cell and that from the remainder of the cell in the forward direction ($\sin\theta/\lambda = 0$) must be identically equal to zero. The electron-density can be considered to be uniformly distributed at low resolution and the scattering angle is close to zero. Therefore, the contributions to the structure factors of reflections at that resolution from the protein and the solvent are almost equal in magnitude but differ in phase by about 180°. This observation is borne out in the data from 2Zn insulin crystals as can be seen from Table 1 which lists the observed structure factors of a few low angle reflections as well as a set of calculated structure factors (for each reflection) which differ only in the number of water molecules included in the calculations. Obviously, the water molecules need not necessarily be accurately placed to produce reasonable agreement between observed and calculated structure factors at this resolution; it is only the overall features of the solvent structure that are important in this respect.

Most workers choose to omit the inner reflections from refinement calculations. It is not perhaps desirable to omit them altogether, especially when the water structure is also sought to be determined, as they contain useful information regarding the general features of solvent distribution. It should however be remembered that even when these reflections are acentric, the relationship between $\bar{F}_P$ and $\bar{F}_Q$ in each of them is that appropriate for centric reflections. The non-random nature of the orientation between $\bar{F}_P$ and $\bar{F}_Q$ leads to larger differences between the magnitudes of the observed and the calculated structure factors than could be expected when the orientations between $\bar{F}_P$ and $\bar{F}_Q$ are randomly

Table 1. Structure factors of a few typical low angle reflections at different stages of 2Zn insulin refinement

h	k	l	F_o	a_{best}	f.m.	Protein		Protein+170H_2O		Protein+200H_2O		Protein+262H_2O	
						F_c	a_c	F_c	a_c	F_c	a_c	F_c	a_c
0	3	0	330	-148	1.0	1978	-175	1127	-173	870	-165	359	-166
-1	4	1	699	- 21	0.99	1235	- 29	952	- 18	859	- 16	668	- 28
-1	1	1	84	- 32	0.92	1460	- 3	875	3	791	11	419	0
-2	3	1	53	- 7	0.81	301	- 22	237	- 25	331	- 28	69	14
-3	4	2	1033	162	1.00	1329	177	1196	172	1206	176	1185	178
-6	2	1	550	171	0.99	987	177	705	177	632	-178	637	178

distributed. Therefore, the Fourier coefficients for these reflections should perhaps be given lower weights when computing maps with (21), (34) or (38) as coefficients. This is automatically achieved in (34) and (38) when the parameters m, n, m' and n' are calculated using empirically evaluated S_P and S_N.

Reliability of results and checking procedures

When good data at atomic resolution are available, as is normally the case with small structures, automatic refinement procedures, by and large, lead to reliable results. The refinement of protein structures is complicated by several factors, the chief among which are the limited nature of the data set, and the size and complexity of the protein. Also, the definition of the structure, and hence the accuracy of the results, vary considerably from one region of the structure to another depending on the flexibility of the group involved. Therefore, the reliability of the refined parameters need to be considered carefully.

Many authors have attempted to estimate errors by theoretical means, for example, using Luzzati diagram (Luzzati, 1952) or Cruickshank's equations (Cruickshank, 1949). Errors have also been estimated from the population variance in bond lengths before regularisation. Comparison of the refined dimensions of chemically equivalent but crystallographically independent molecules in the same crystal or different crystals is yet another method employed for estimating errors. In this method, however, the observed differences contain contributions from errors as well as genuine differences arising from differences in intermolecular interactions.

Perhaps the most interesting results pertaining to the reliability of refined coordinates are those obtained during the refinement of 2Zn insulin crystals (Dodson et al. 1979). The structure was refined simultaneously by the difference Fourier method in Professor Dorothy Hodgkin's Laboratory at Oxford and by the least squares method (Isaacs and Agarwal, 1978) at IBM Research Center, New York. The starting set of coordinates in the former was obtained from an isomorphously and anomalously phased 1.9 Å map. The starting set of coordinates in the latter was derived from a map computed from phases refined using the method due to Sayre (1972). The same data at 1.5 Å resolution were used in both the refinements. During the course of both the refinements, automatic procedures were often interrupted to add, delete or modify parts of the structure including solvent molecules. Each refinement produced a set of coordinates which was internally consistent and also led to acceptable agreement between observed and calculated structure factors

($R < 0.20$ for the entire data set). When compared, the two sets of coordinates were found to agree within 0.3 Å for a majority of protein atoms. There were, however, larger descrepancies in the position of the remaining protein atoms, which all belonged to surface residues, and water molecules. In most instances, the descrepancies corresponded to differences in detail, but in a few cases they represented gross differences in side chain conformation. Thus, the agreement between the two sets of coordinates, though good for the residues in the interior of the molecule, was on the whole rather disappointing. The descrepancies in the results were subsequently resolved and the correct positions arrived at through a detailed manual examination of different types of Fourier maps, including those in the earlier stages of refinement, coupled with geometrical and chemical considerations.

The experience cited above clearly shows that the usual crystallographic indicators for the convergence of refinement do not assure that the refined parameters, especially those pertaining to ill-defined structure, are necessarily correct. Often an interpretation of diffuse density, even when essentially wrong, does not get automatically corrected during the course of the refinement. This emphasises the need for thorough periodic checks on the current set of coordinates. One of the methods employed for this purpose consists in leaving out part of the structure from the structure factor calculations and then checking the atomic coordinates in that part of the structure against the subsequent difference Fourier map. As an example for this procedure, one set of calculations carried out towards the end of the refinement of 2Zn insulin may be outlined here. In this set of calculations, the asymmetric unit ($0\text{-}a/3,\ 0\text{-}b/3,\ 0\text{-}c$) was divided into eight segments along the c axis. To start with, all the atoms in the first segment (0 to $c/8$) were removed from structure factor calculations and a difference Fourier map was computed covering only this segment. The operation was repeated for the remaining seven segments as well. Compiling the separately computed difference densities in the eight segments, one obtained a map in which the density in any given segment was formally independent of the input atoms in that segment. The abrupt changes in density which, for obvious reasons, occurred at the boundaries between adjacent segments made the interpretation in the neighbourhood of the boundaries rather difficult. Therefore, another set of difference Fourier maps was also computed with segments $-c/16$ to $c/16$, $c/16$ to $3c/16$ etc. The two sets of difference Fourier maps provided a valuable check on the current set of coordinates. A region in one of the difference Fourier maps and the same region in the F_o map are shown in Figure 8(a) and (b) respectively for comparison.

It would appear that the different types of Fourier syntheses are not as effective as one would normally expect in indicating the errors in the

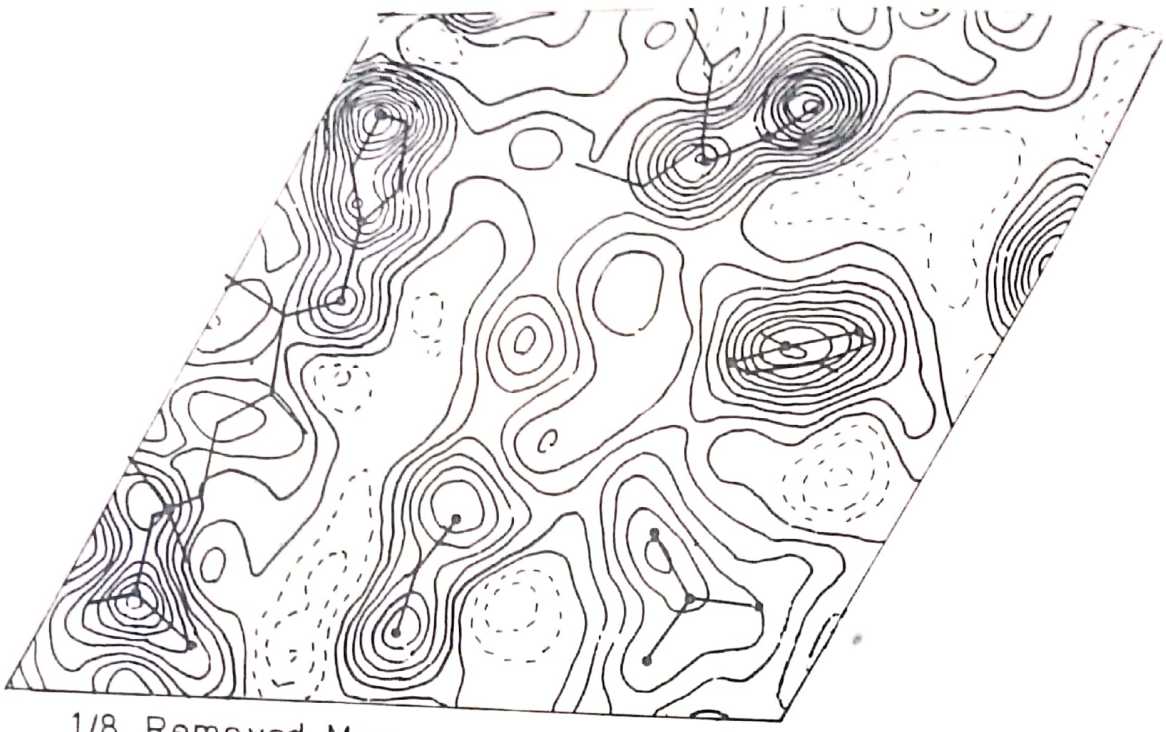


Figure 8 (a)

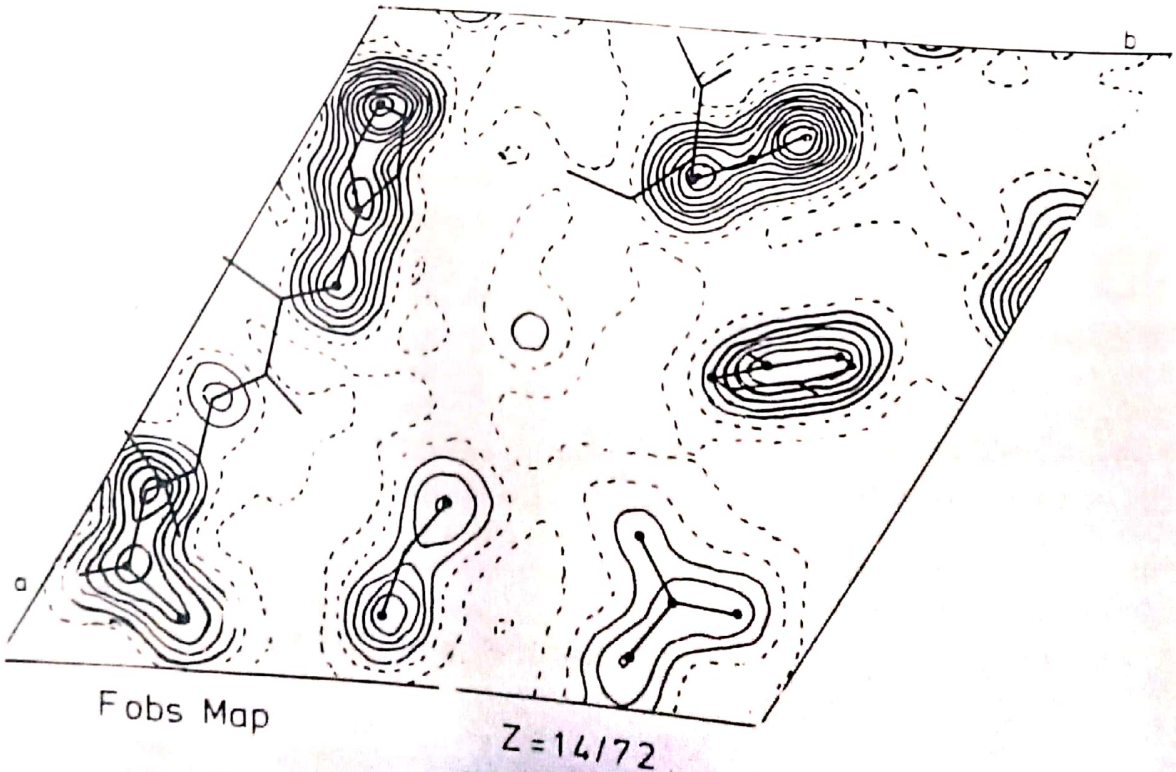


Figure 8 (b)

positions of the known atoms and in giving the correct positions of the unknown atoms. When reconciling the two sets of coordinates in 2Zn insulin, arrived at through different refinement procedures, it appeared that some atoms "remembered" their previous history even when they were not included in the calculation of phase angles. One explanation for this phenomenon might be related to the effect of errors in the positions of the input atoms on the corrected positions of the known atoms and the positions of the unknown atoms obtained from Fourier syntheses. It may be recalled that (31) to (33) were derived from (25) to (30) on the assumption that the errors in $\bar{r}_{Pj}$ were small and random. It can be readily shown that non-random errors lead to shifts in peak positions from $\bar{r}_{Pj}$ and $\bar{r}_{Qj}$. When the errors are systematic and large, features in Fourier maps can no longer be divided in a simple manner into those contributing to peaks at atomic positions and those contributing to background. What was considered earlier as background is also then likely to be important in determining peak positions. It is also clear from the analysis that there is no inherent mechanism for automatically correcting the positional errors (in Fourier maps) resulting from non-random errors in the positions of the input atoms. It is therefore important to carry out careful manual examination of different Fourier maps at various stages of refinement to make sure that large systematic errors are not introduced in the input coordinates.

The author thanks the University Grants Commission, India, for financial assistance.

References

Adams, M.J. (1968). Ph.D. Thesis, Oxford University.
Ashida, T. (1976). In Crystallographic Computing Techniques, p. 282, ed. F.R. Ahmed, Copenhagen: Munksgaard.
Blake, C.C.F., Geisow, M.J. and Oatley, S.J. (1978). J. Mol. Biol. 121, 339.
Blow, D.M. and Crick, F.H.C. (1959). Acta Cryst. 12, 794.
Bode, W. and Schwager, P. (1975). J. Mol. Biol. 98, 693.
Chambers, J.L. and Stroud, M. (1977). Acta Cryst. B33, 1824.
Cruickshank, D.W.J. (1949). Acta Cryst. 2, 65.
Cullis, A.F., Muirhead, H., Perutz, M.F., Rossmann, M.G. and North, A.C.T. (1961). Proc. Roy. Soc. A265, 15.
Deisenhofer, J. and Steigemann, W. (1975). Act Cryst. B31, 238.
Diamond, R. (1966). Acta Cryst. 21, 253.
Diamond, R. (1971). Acta Cryst. A27, 435.
Dickerson, R.E., Kendrew, J.C. and Strandberg, B.E. (1961). Acta Cryst. 14, 1188.

Dodson, E.J., Dodson, G.G., Hodgkin, D.C. and Vijayan, M. (1979). Unpublished results.
Dodson, E.J., Isaacs, N.W. and Rollett, J.S. (1976). Acta Cryst. A32, 311.
Dodson, E.J. and Vijayan, M. (1971). Acta Cryst. B27, 2402.
Einstein, J.R. (1977). Acta Cryst. A33, 75.
Epp, O., Lattman, E.E., Schiffer, M., Huber, R. and Palm, W. (1975). Biochemistry 14, 4943.
Freer, S.T., Alden, R.A., Carter, Jr., C.W. and Kraut, J. (1975). J. Biol. Chem. 250, 46.
Green, E.A. (1979). Acta Cryst. A35, 351.
Harker, D. (1956). Acta Cryst. 9, 1.
Hendrickson, W.E. and Lattman, E.E. (1970). Acta Cryst. B36, 136.
Hermans, J. and McQueen, J.E. (1974). Acta Cryst. A30, 730.
Huber, R., Kukla, D., Bode, W., Schwager, P., Bartels, K., Deisenhofer, J. and Steigemann, W. (1974). J. Mol. Biol. 89, 73.
International Tables for X-ray Crystallography (1962). Vol. III. Birmingham: Kynoch Press.
Isaacs, N.W. and Agarwal, R.C. (1978). A34, 782.
Kartha, G. (1976). In Crystallographic Computing Techniques, p. 269, ed. F.R. Ahmed, Copenhagen: Munksgaard.
Kartha, G. and Parthasarathy, R. (1965). Acta Cryst. 18, 745.
Luzzati, V. (1952). Acta Cryst. 5, 802.
Luzzati, V. (1953). Acta Cryst. 6, 142.
Matthews, B.W. (1966). Acta Cryst. 20, 230.
Moews, P.C. and Kretsinger, R.H. (1975). J. Mol. Biol. 91, 201.
North, A.C.T. (1965). Acta Cryst. 18, 212.
Raiz, V.Sh. and Andreeva, N.S. (1970). Sov. Phys. Crystallogr. 15, 210.
Ramachandran, G.N. and Raman, S. (1956). Curr. Sci. 25, 348.
Ramachandran, G.N. and Srinivasan, R. (1970). Fourier Methods in Crystallography. New York: John Wiley and Sons, Inc.
Sayre, D. (1972). Acta Cryst. A28, 210.
Singh, A.K. and Ramaseshan, S. (1966). Acta Cryst. 21, 279.
Stenkamp, R.E., Sieker, L.C. and Jensen, L.H. (1978). Acta Cryst. A34, 1014.
Vijayan, M. (1980). Acta Cryst. A36, 295.
Watenpaugh, K.D., Sieker, L.C., Herriot, J.R. and Jensen, L.H. (1973). Acta Cryst. B29, 943.

EXERCISES

Problems

I. Determine a_p for reflections with the following set of parameters using Harker diagrams.

1) $F_p = 55$,
$F_{H1} = 20$, $a_{H1} = 84°$, $F_{PH1} = 67$,
$FH_2 = 30$, $a_{H2} = -14.5°$, $F_{PH2} = 82$,

2) $F_p = 50$,
$F_H = 20$, $a_H = 107°$, $F''_H = 2.5$,
$F^+_{PH} = F^-_{PH} = 70$

3) $F_p = 50$,
$F_H = 20$, $a_H = 197°$, $F''_H = 2.5$,
$F^+_{PH} = 51.5$, $F^-_{PH} = 56$

II. Compute m and a_B for the following reflections using eqns. (2) to (6) in the text.

1) $F_p = 54$, $F_{H1} = 19$, $a_{H1} = 80°$, $F_{PH1} = 68$
$FH2 = 32$, $a_{H2} = -12°$, $F_{PH2} = 77$

a) $E_1 = 10$, $E_2 = 20$; (b) $E_1 = 4$, $E_2 = 8$

2) $F_p = 56$, $F_{H1} = 20$, $a_{H1} = 95°$, $F_{PH1} = 70$
$F_{H2} = 20$, $a_{H2} = -90°$, $F_{PH2} = 75$

a) $E_1 = 10$, $E_2 = 20$; (b) $E_1 = 4$, $E_2 = 8$

Answers

I. 1. $a_p = 21°$
2. $a_p = 107°$
3. $a_p = 107°$

II. 1. (a). $m = 0.695$, $a_B = 45.4°$
(b). $m = 0.952$, $a_B = 38.0°$

2. (a). $m = 0.484$, $a_B = 159.1°$
(b). $m = 0.548$, $a_B = 154.8°$